

# Multiple Sequence Alignment Theory & Practice

---

*Peter FitzGerald & Susan Chacko*

NCI & CIT

# Outline

---

- Introduction to MSA
  - What is it ?
  - What is it good for ?
  - How do I use it ?
- Software and algorithms
  - The programs
  - How they work
  - Which to use
  - Editing & publishing
- Conclusion & Recommendations
- Multiple Genome Alignment

# What is Multiple Sequence Alignment (MSA)?

---

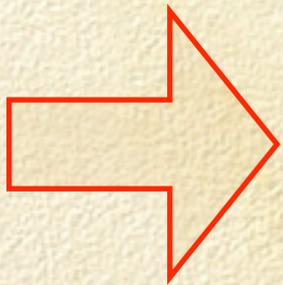
<b>chicken</b>	P	L	V	S	S	-	-	-	P	L	R	G	E	A	G	V	L	P	F	Q	Q	E	E	Y	E	K	V	K	R	G	I	V	E	Q	C	C	H	N	T	C	S	L	Y	Q	L	E	N	Y	C	N
<b>xenopus</b>	A	L	V	S	G	-	-	-	P	Q	D	N	E	L	D	G	M	Q	L	Q	P	Q	E	Y	Q	K	M	K	R	G	I	V	E	Q	C	C	H	S	T	C	S	L	F	Q	L	E	S	Y	C	N
<b>human</b>	L	Q	V	G	Q	V	E	L	G	G	G	P	G	A	G	S	L	Q	P	L	A	L	E	G	S	L	Q	K	R	G	I	V	E	Q	C	C	T	S	I	C	S	L	Y	Q	L	E	N	Y	C	N
<b>monkey</b>	P	Q	V	G	Q	V	E	L	G	G	G	P	G	A	G	S	L	Q	P	L	A	L	E	G	S	L	Q	K	R	G	I	V	E	Q	C	C	T	S	I	C	S	L	Y	Q	L	E	N	Y	C	N
<b>dog</b>	L	Q	V	R	D	V	E	L	A	G	A	P	G	E	G	L	Q	P	L	A	L	E	G	A	L	Q	K	R	G	I	V	E	Q	C	C	T	S	I	C	S	L	Y	Q	L	E	N	Y	C	N	
<b>hamster</b>	P	Q	V	A	Q	L	E	L	G	G	G	P	G	A	D	D	L	Q	T	L	A	L	E	V	A	Q	Q	K	R	G	I	V	D	Q	C	C	T	S	I	C	S	L	Y	Q	L	E	N	Y	C	N
<b>cow</b>	P	Q	V	G	A	L	E	L	A	G	G	P	G	A	G	G	-	-	-	-	-	L	E	G	P	P	Q	K	R	G	I	V	E	Q	C	C	A	S	V	C	S	L	Y	Q	L	E	N	Y	C	N
<b>guinea pig</b>	P	Q	V	E	Q	T	E	L	G	M	G	L	G	A	G	G	L	Q	P	L	A	L	E	M	A	L	Q	K	R	G	I	V	D	Q	C	C	T	G	T	C	T	R	H	Q	L	Q	S	Y	C	N
	*																					*					*	*	*	*	*	:	*	*	*		*	:		*	*	:	.	*	*	*		*	*	*

# Why do a Multiple Sequence Alignment ?

What's the end goal ?

---

- Simple sequence comparison
- Conserved *vs.* non-conserved regions
  - proteins - motifs/profiles
  - whole genome - genes, control regions
- Homology (as opposed to similarity)
  - Evolution - phylogeny
  - Structural homology
- Sequence differences
  - Single Nucleotide Polymorphisms (SNPs)



# Subsets of Functions

---

- ❑ Multiple Alignment
- ❑ Multiple Sequence Editing
- ❑ Generating/drawing trees
- ❑ Publishing - high quality output
- ❑ Structure interface (CN<sub>3</sub>D)

# Pre-computed MSAs

---

- DALI/FSSP  
<http://www2.ebi.ac.uk/dali/>
- InterPro  
<http://www.ebi.ac.uk/interpro/>
- PROSITE, PRINTS  
<http://us.expasy.org/prosite/>
- CDD, SMART, PFAM, COG  
<http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>
- VAST  
<http://www.ncbi.nlm.nih.gov:80/Structure/VAST/vastsearch.html>

# Domain/Profile Construction

---

- ❑ PSI-BLAST  
*<http://www.ncbi.nlm.nih.gov/BLAST/>*
- ❑ MEME/MAST  
*<http://meme.sdsc.edu/meme/website/intro.html>*
- ❑ BLOCKS  
*<http://www.blocks.fhcrc.org/>*
- ❑ PRATT  
*<http://us.expasy.org/tools/pratt/>*
- ❑ HMMER  
*<http://hmmerr.wustl.edu/>*

# Generating an Alignment

---

- Get the sequences
  - ◆ Reformat them
- Align the sequences
  - ◆ Evaluate the alignment
  - ◆ Realign or modify the alignment
  - ◆ Add/subtract sequence
- Analyze, publish, draw phylogenetic trees, connect to structures

# Collecting the Sequences

---

- Selection of sequences is important
  - Most programs will align ***ANYTHING***
  - All sequences should be related
  - Avoid redundancy
  - Diverse set of sequences is best

# Sequence Selection

---

- ❑ Common source of sequences is blast output
- ❑ Entrez searches
- ❑ Many pre-aligned
- ❑ Personal sequences

# Sequence Format

---

- ❑ Several multiple sequence formats
- ❑ Format selection is important for input and output
- ❑ Different programs *like* (**need**) different formats
- ❑ Reformatting software  
*<http://molbio.info.nih.gov/molbio/gcglite/reformat.html>*  
*<http://genome.nci.nih.gov/tool/reformat.html>*
- ❑ Output format determined by next step

# Sequence formats

*(sequential)*

---

>**chiins** *insulin2.msf*, 107 aa.

```
BALWIRSLPLLALLVFSGPGTSYAAANQHLCGSHLVEALYLVCGERGFFYSPKARRDVEQ  
PLVSS---PLRGEAGVLPFQQEYKVKRGIVEQCCHNTCSLYQLENYCN
```

>**xenins** *insulin2.msf*, 106 aa.

```
BALWMQCLPLVLVLFSTPNT-ALVNQHLCGSHLVEALYLVCGDRGFFYYPKVKRDMEQ  
ALVSG---PQDNELDGMQLQPQEQKMKRGIVEQCCHSTCSLFLQLESYCN
```

>**humins** *insulin2.msf*, 110 aa.

```
BALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED  
LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
```

>**monins** *insulin2.msf*, 110 aa.

```
BALWMRLLPLLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED  
PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
```

>**dogins** *insulin2.msf*, 110 aa.

```
MALWMRLLPLLALLALWAPAPTRAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVED  
LQVRDVELAGAPGEGGLQPLALEGALQKRGIVEQCCTSICSLYQLENYCN
```

>**hamins** *insulin2.msf*, 110 aa.

```
MTLWMRLLPLLTLVLWEPNPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKSRRGVED  
PQVAQLELGGGPGADDLQTLALEVAQQKRGIVDQCCTSICSLYQLENYCN
```

>**bovins** *insulin2.msf*, 105 aa.

```
MALWTRLRPLLALLALWPPPPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEG  
PQVGALELAGGPGAGG-----LEGPPQKRGIVEQCCASVCSLYQLENYCN
```

>**guins** *insulin2.msf*, 110 aa.

```
MALWMHLLTVLALLALWGPNTGQAFVSRHLCGSNLVETLYSVCQDDGFFYIPKDRRELED  
PQVEQTELGMGLGAGGLQPLALEMALQKRGIVDQCCTGTCTRHLQLSYCN
```

# ClustalW

*(interlaced)*

---

**CLUSTAL W (1.74) multiple sequence alignment**

<b>chiins</b>	BALWIRSLPLLALLVFSGPGTSYAAANQHLCGSHLVEALYLVCGERGFFYSPKARRDVEQ
<b>xenins</b>	BALWMQCLPLVLVLFSTPNT-ALVNQHLCGSHLVEALYLVCGERGFFYYPKVKRDMEQ
<b>humins</b>	BALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
<b>monins</b>	BALWMRLLPLLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
<b>dogins</b>	MALWMRLLPLLALLALWAPAPTRAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVED
<b>hamins</b>	MTLWMRLLPLLTLVLEWPNPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKSRRGVED
<b>bovins</b>	MALWTRLRPLLALLALWPPPPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEG
<b>guiins</b>	MALWMHLLTVLALLALWGPNTGQAFVSRHLCGSNLVETLYSVCQDDGFFYIPKDRRELED

<b>chiins</b>	PLVSS---PLRGEAGVLPFQQEYKVKRGIVEQCCHNTCSLYQLENYCN
<b>xenins</b>	ALVSG---PQDNELDGMQLQPQEQKMKRGIVEQCCHSTCSLFOLESYCN
<b>humins</b>	LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
<b>monins</b>	PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
<b>dogins</b>	LQVRDVELAGAPGEGGLQPLALEGALQKRGIVEQCCTSICSLYQLENYCN
<b>hamins</b>	PQVAQLELGGGPGADDLQTLALEVAQQKRGIVDQCCTSICSLYQLENYCN
<b>bovins</b>	PQVGALELAGGPGAGG-----LEGPPQKRGIVEQCCASVCSLYQLENYCN
<b>guiins</b>	PQVEQTELGMLGAGGLQPLALEMALQKRGIVDQCCTGTCTRHLQLSYCN

# GCG - MSF - Pileup

## PileUp

MSF: 110 Type: P Check: 4380 ..

Name: chiins oo Len: 110 Check: 3857 Weight: 0.212  
Name: xenins oo Len: 110 Check: 4552 Weight: 0.050  
Name: humins oo Len: 110 Check: 4867 Weight: 0.050  
Name: monins oo Len: 110 Check: 5690 Weight: 0.080  
Name: dogins oo Len: 110 Check: 3667 Weight: 0.111  
Name: hamins oo Len: 110 Check: 5715 Weight: 0.111  
Name: bovins oo Len: 110 Check: 845 Weight: 0.232  
Name: guins oo Len: 110 Check: 5187 Weight: 0.100

//

<b>chiins</b>	BALWIRSLPL	LALLVFSGPG	TSYAAANQHL	CGSHLVEALY	LVCGERGFFY
<b>xenins</b>	BALWMQCLPL	VLVLFSTPN	TE.ALVNQHL	CGSHLVEALY	LVCGDRGFFY
<b>humins</b>	BALWMRLLPL	LALLALWGP	PAAAFVNQHL	CGSHLVEALY	LVCGERGFFY
<b>monins</b>	BALWMRLLPL	LALLALWGP	PVPAFVNQHL	CGSHLVEALY	LVCGERGFFY
<b>dogins</b>	MALWMRLLPL	LALLALWAPA	PTRAFVNQHL	CGSHLVEALY	LVCGERGFFY
<b>hamins</b>	MTLWMRLLPL	LTLLVLWEPN	PAQAFVNQHL	CGSHLVEALY	LVCGERGFFY
<b>bovins</b>	MALWTRLRPL	LALLALWPPP	PARAFVNQHL	CGSHLVEALY	LVCGERGFFY
<b>guins</b>	MALWMHLLTV	LALLALWGPN	TGQAFVSRHL	CGSNLVETLY	SVCQDDGFFY

<b>chiins</b>	SPKARRDVEQ	PLVSS...PL	RGEAGVLPFQ	QEEYEKVKRG	IVEQCCHNTC
<b>xenins</b>	YPKVKRDMEQ	ALVSG...PQ	DNELDGMQLQ	PQEYQMKRG	IVEQCCHSTC
<b>humins</b>	TPKTRREAED	LQVGQVELGG	GPGAGSLQPL	ALEGSLOKRG	IVEQCCTSIC
<b>monins</b>	TPKTRREAED	PQVGQVELGG	GPGAGSLQPL	ALEGSLOKRG	IVEQCCTSIC
<b>dogins</b>	TPKARREVED	LQVRDVELAG	APGEGGLQPL	ALEGALQKRG	IVEQCCTSIC
<b>hamins</b>	TPKSRRGVED	PQVAQLELGG	GPGADDLQTL	ALEVAQQKRG	IVDQCCTSIC
<b>bovins</b>	TPKARREVEG	PQVGALELAG	GPGAGG....	.LEGPPQKRG	IVEQCCASVC
<b>guins</b>	IPKDRRELED	PQVEQTELM	GLGAGGLQPL	ALEMALQKRG	IVDQCCTGTC

# Generating the Alignment

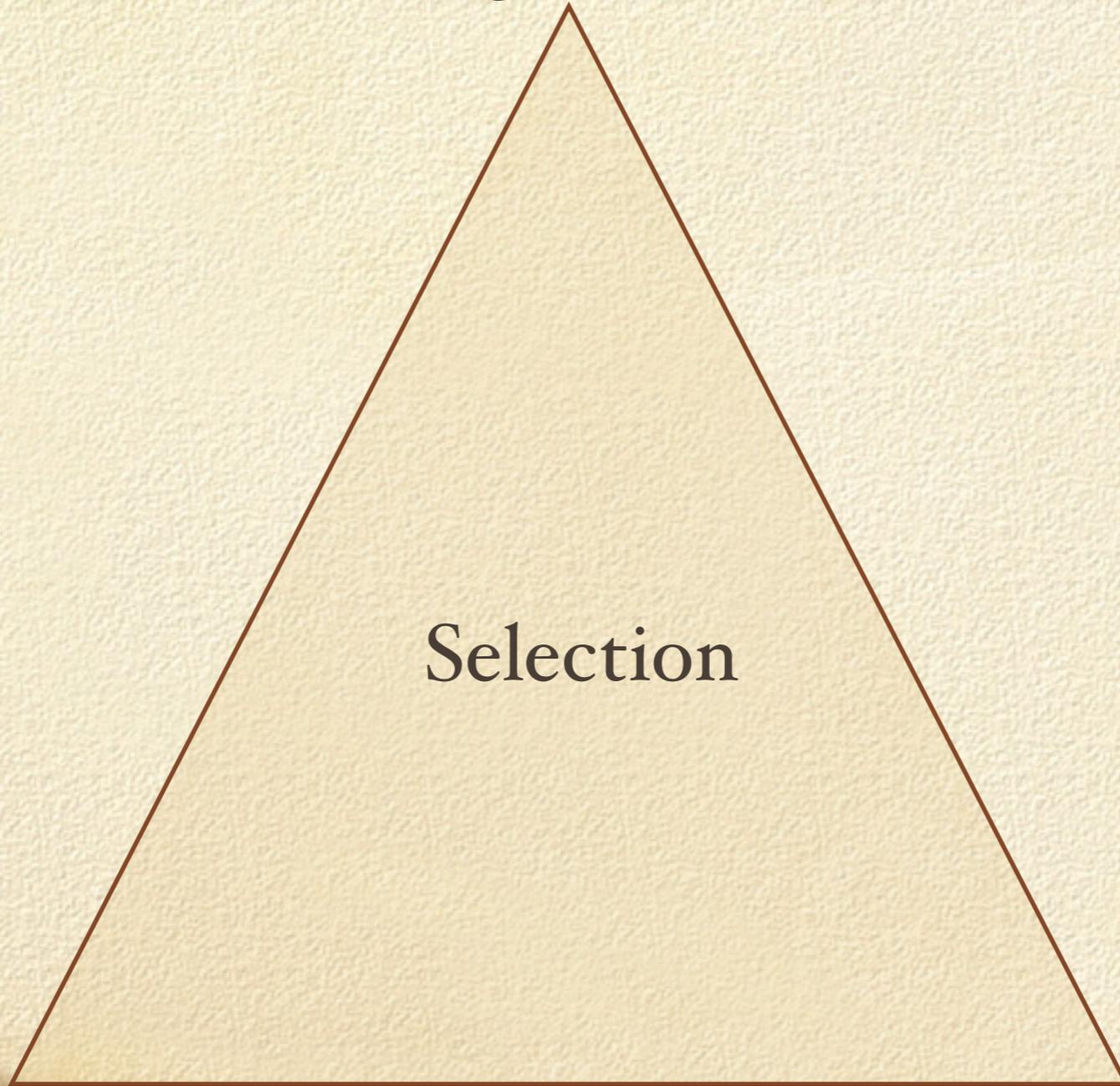
---

Algorithm

Selection

Software

Platform



# Platform - selection

*Choice dependent on availability, complexity and personal preference*

---



Central server



Web-based



Local computer

# Software - selection

*Choice dependent on ease of use and availability*

---

- ❑ The best
- ❑ What's available
- ❑ The easiest to use
- ❑ The best output

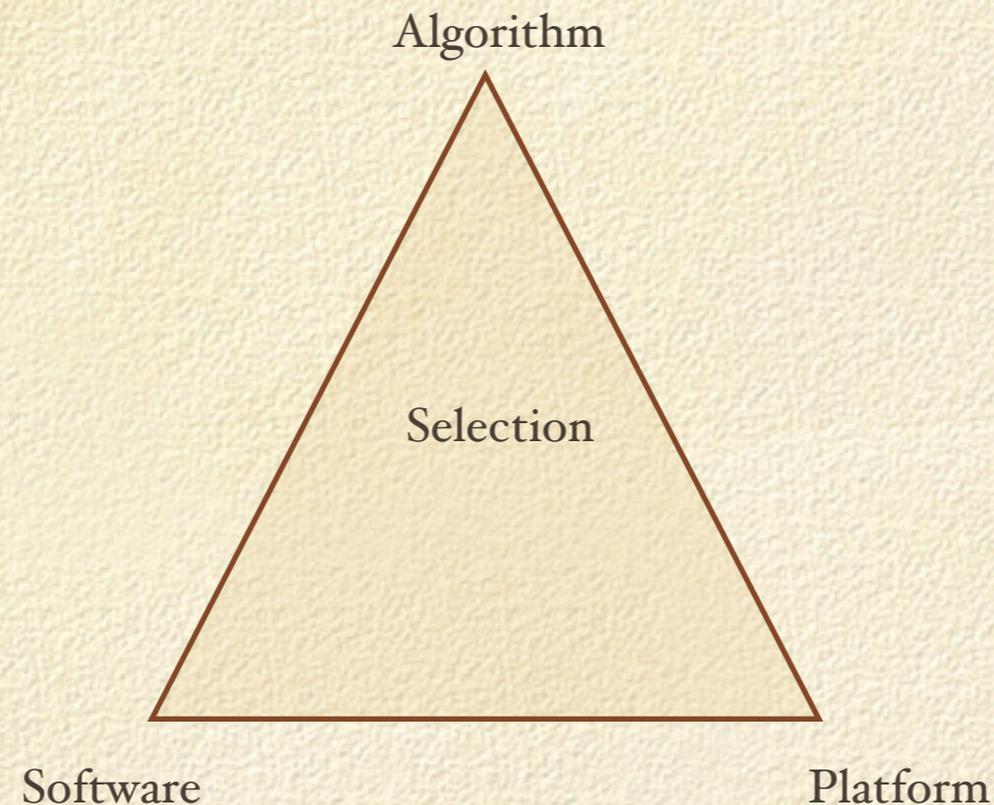
# Algorithm - selection

---

- ❑ The most accurate
- ❑ The best for your problem
- ❑ What's available
- ❑ What you are familiar with

# Generating the Alignment

---



*Natural selection* of software - driven by ease of use and availability (portability) - determines which programs are used most frequently.

# MSA Programs

*(a sampling)*

---

Allall	Blast	Blocks
DiAlign	Dalign	DCA
Dali	Clustalw	ClustalX
ComAlign	GA	HMMER
IterAlign	MAVID	MAFFT
MSA	MultAlign	MultAlin
Musca	Museqal	Oma
T-Coffee	ToPLign	TreeAlign
Pileup(GCG)	POA	Praline
PRRP	SAM	SAGA

# MSA Programs

*(the focus)*

---

- MSA

(close-to-) optimal Alignments using the Carrillo-Lipman bound

- ClustalW/ClustalX

the most widely used program for multiple alignment

- T-Coffee

allows the combination of a collection of multiple/pairwise, global or local alignments into a single model

- DiAlign

constructs pairwise and multiple alignments by comparing whole *segments* of the sequences. No gap penalty is used

- POA & MAFFT

POA: partial order alignment, based on a graph representation of an MSA

MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform

Name	Algorithm	URL
MSA	<b>EXACT</b>	<a href="http://www.ibc.wustl.edu/ibc/msa.html">http://www.ibc.wustl.edu/ibc/msa.html</a>
DCA (requires MSA)	Exact	<a href="http://bibiserv.techfak.uni-bielefeld.de/dca">http://bibiserv.techfak.uni-bielefeld.de/dca</a>
OMA	Iterative DCA	<a href="http://bibiserv.techfak.uni-bielefeld.de/oma">http://bibiserv.techfak.uni-bielefeld.de/oma</a>
ClustalW, ClustalX	<b>PROGRESSIVE</b>	<a href="ftp://ftp-igbmc.u-strasbg.fr/pub/clustalW">ftp://ftp-igbmc.u-strasbg.fr/pub/clustalW</a> or <a href="ftp://ftp-igbmc.u-strasbg.fr/pub/clustalX">clustalX</a>
MultAlign	Progressive	<a href="http://www.toulouse.inra.fr/multalin.html">http://www.toulouse.inra.fr/multalin.html</a>
Dialign	<b>CONSISTENCY BASED</b>	<a href="http://www.gsf.de/biodv/dialign.html">http://www.gsf.de/biodv/dialign.html</a>
ComAlign	Consistency-based	<a href="http://www.daimi.au.dk/~ocaprani">http://www.daimi.au.dk/~ocaprani</a>
T-Coffee	<b>CONSISTENCY BASED/PROGRESSIVE</b>	<a href="http://igs-server.cnrs-mrs.fr/~cnotred">http://igs-server.cnrs-mrs.fr/~cnotred</a>
Praline Iterative/progressive	Iterative/ progressive	<a href="mailto:jhering@nimr.mrc.ac.uk">jhering@nimr.mrc.ac.uk</a>
IterAlign Iterative	Iterative	<a href="http://giotto.Stanford.edu/~luciano/iteralign.html">http://giotto.Stanford.edu/~luciano/iteralign.html</a>
Prrp	Iterative/Stochastic	<a href="ftp://ftp.genome.ad.jp/pub/genome/saitama-cc/">ftp://ftp.genome.ad.jp/pub/genome/saitama-cc/</a>
SAM	Iterative/ Stochastic/HMM	<a href="mailto:rph@cse.ucsc.edu">rph@cse.ucsc.edu</a>
HMMER	Iterative/ Stochastic/HMM	<a href="http://hmmer.wustl.edu/">http://hmmer.wustl.edu/</a>
SAGA	Iterative/ Stochastic/GA	<a href="http://igs-server.cnrs-mrs.fr/~cnotred">http://igs-server.cnrs-mrs.fr/~cnotred</a>
GA	Iterative/ Stochastic/GA	<a href="mailto:czhang@watnow.uwaterloo.ca">czhang@watnow.uwaterloo.ca</a>

# Multiple Sequence Alignment Methods

---

- Local Alignment-----Global Alignment
- Exact (MSA, DCA)  
*good for few, short, closely related sequences*
- Progressive alignment (ClustalW)  
*fast, sensitive*
- Consistency based method (DiAlign)  
*better for sequences with large insertions*
- Iterative method (HMMER, SAM, HMMs)  
*slow, sometimes inaccurate ...good for profiles*
- Combination methods (T-coffee)  
*very good but can be slow*

# Aligning two sequences

*(Needleman and Wunsch)*

	<b>F</b>	<b>A</b>	<b>S</b>	<b>T</b>
<b>F</b>	•			
<b>A</b>		•		
<b>T</b>				•

*10 sequences (100AA)*

*3 million years and 10 billion gigabytes*

Match=1

Mismatch= -1

Gap = -1

**FAST**

**FA-T**

		<b>F</b>	<b>A</b>	<b>S</b>	<b>T</b>
	0	-1	-2	-3	-4
<b>F</b>	-1	1	0		
<b>A</b>	-2	0	2		
<b>T</b>	-3				

		<b>F</b>	<b>A</b>	<b>S</b>	<b>T</b>
	0	-1	-2	-3	-4
<b>F</b>	-1	1	0		
<b>A</b>	-2	0	2	1	0
<b>T</b>	-3			1	2

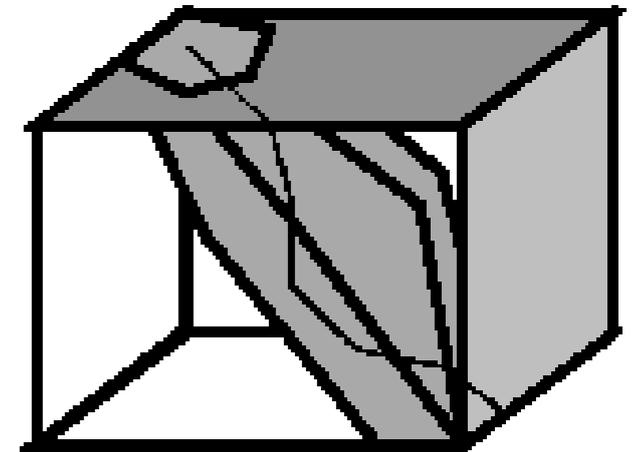
		<b>F</b>	<b>A</b>	<b>S</b>	<b>T</b>
	0				
<b>F</b>		1			
<b>A</b>			2	1	
<b>T</b>					2

# MSA

## *An EXACT Alignment*

---

1. Determine the optimal **pairwise alignments**.
2. Perform a **fast multiple sequence alignment** (progressive) and extract the pairwise alignments from this multiple sequence alignment.
3. For each pair of sequences, use the optimal and extracted pairwise alignments to define the **restricted alignment space** defined by the difference in the two alignment scores for this pair of sequences.
4. Project the restricted pairwise alignment spaces into the multidimensional alignment space to define the **restricted hyper-volume of the multidimensional space** to determine the best multiple sequence alignment. The greater the overall sequence similarity, the smaller the restricted alignment space is.
5. Use dynamic programming to compute the **value for all of the cells** within the restricted alignment space.
6. **Backtrack through the restricted alignment space** to recover the best alignment. The result is a minimum distance alignment.



# MSA - 4GB memory

*limited by length and diversity of sequence*

---

- 20 phospholipase (130 AA)
- 14 (highly diverse) cytochrome C (110)
- 6 (moderately diverse) aspartyl proteases (350)
- 8 (moderately diverse) lipid-binding proteins(480)

# ClustalW

## *A Progressive Alignment*

---

- 1. Pairwise Distances** - Perform Needleman-Wunsch (global) alignment on all sequence pairs to find the distance between all pairs of sequences.
- 2. Cluster the Pairwise Distances** - Perform a simple clustering to determine which pairs of sequences are closer than others. Using pairwise alignments iteratively one can create phylogenetic relationships, which then allows for the creation of either a UPGMA-constructed guide tree or a Neighbor-Joining guide tree (both rooted trees). These joining trees are based on alignment scores and non-biological rules for creating trees; thus, they should be used cautiously as an evolutionary tree. This step represents a major difference among the various implementations of the PPA and is the part of the algorithm where some of the greatest improvements have occurred.
- 3. Align the Sequences Guided by Clustering** - Align the closest sequences in the joining tree together, followed by adding more sequences to the the initial alignment. For example, when using an UPGMA guide tree or Neighbor-Joining guide tree, one would align a pair of sequences by starting at the bottom of a branch and successively adding more sequences to the nascent alignment (the nascent alignment defines the range of possibilities for the ancestral sequence).

# ClustalW issues

---

- Choice of input sequences
- Order of sequences in (tree)
- Parameters
  - weighting, substitution matrix, gap penalties*
- Progressive (*once a gap always a gap*)
- Known to miss some conserved residues

# T-Coffee

*allows the combination of a collection of multiple/  
pairwise, global or local alignments into a single model*

- ❑ Pairwise global alignment
- ❑ Pairwise local alignment
- ❑ Combined above two into a library
- ❑ Builds MSA with highest consistency with the library of alignments (progressive assembly)

# DiAlign

*constructs pairwise and multiple alignments by comparing whole segments of the sequences.*

---

- Alignment of whole segments and not individual amino acids (bases)
- Pair wise comparison > segment pairs (diagonals), *represent local alignments*
- Diagonals weighted for likelihood
- Alignment built from consistent diagonals
- No gap penalties
- Independent of sequence order

# Meaningfulness

---

- ❑ Is the alignment *correct* ?
- ❑ Can I make it *better* ?
- ❑ Which programs are *best* ?
- ❑ How do you *know* if its correct ?

# Is the Alignment *Correct* ?

---

- What do mean by correct ?
  - Mathematically rigorous
  - Biologically meaningful
  - Operationally useful

# Can you make it *better* ?

---

- ❑ Only if you know what you doing !
- ❑ Define better ?
- ❑ What's the goal ?
- ❑ What's the biology ?

# Which programs are *best* ?

- ❑ No simple answer
- ❑ Depends on the particular problem
- ❑ Recent objective studies help answer this problem
- ❑ Some tools to help compare alignments

# How do you *know* it is correct ?

---

- ❑ Methods to evaluate the alignment
- ❑ Methods to evaluate the program/algorithm
- ❑ Structural information
- ❑ Biology

# Systematic Comparison of MSA programs

---

- **BAlIbASE: a benchmark alignment database for the evaluation of multiple alignment programs**

*Thompson JD, Plewniak F, Poch O. Bioinformatics. 1999 Jan;15(1):87-8.*

- **A comprehensive comparison of multiple sequence alignment programs**

*JD Thompson, F Plewniak, and O Poch Nucleic Acids Res. 1999 27: 2682-2690.*

- **Quality assessment of multiple alignment programs**

*FEBS Letters Volume 529, Issue 1 , T. Lassmann and E Sonnhammer 2 October 2002, Pages 126-130*

# BALiBase -

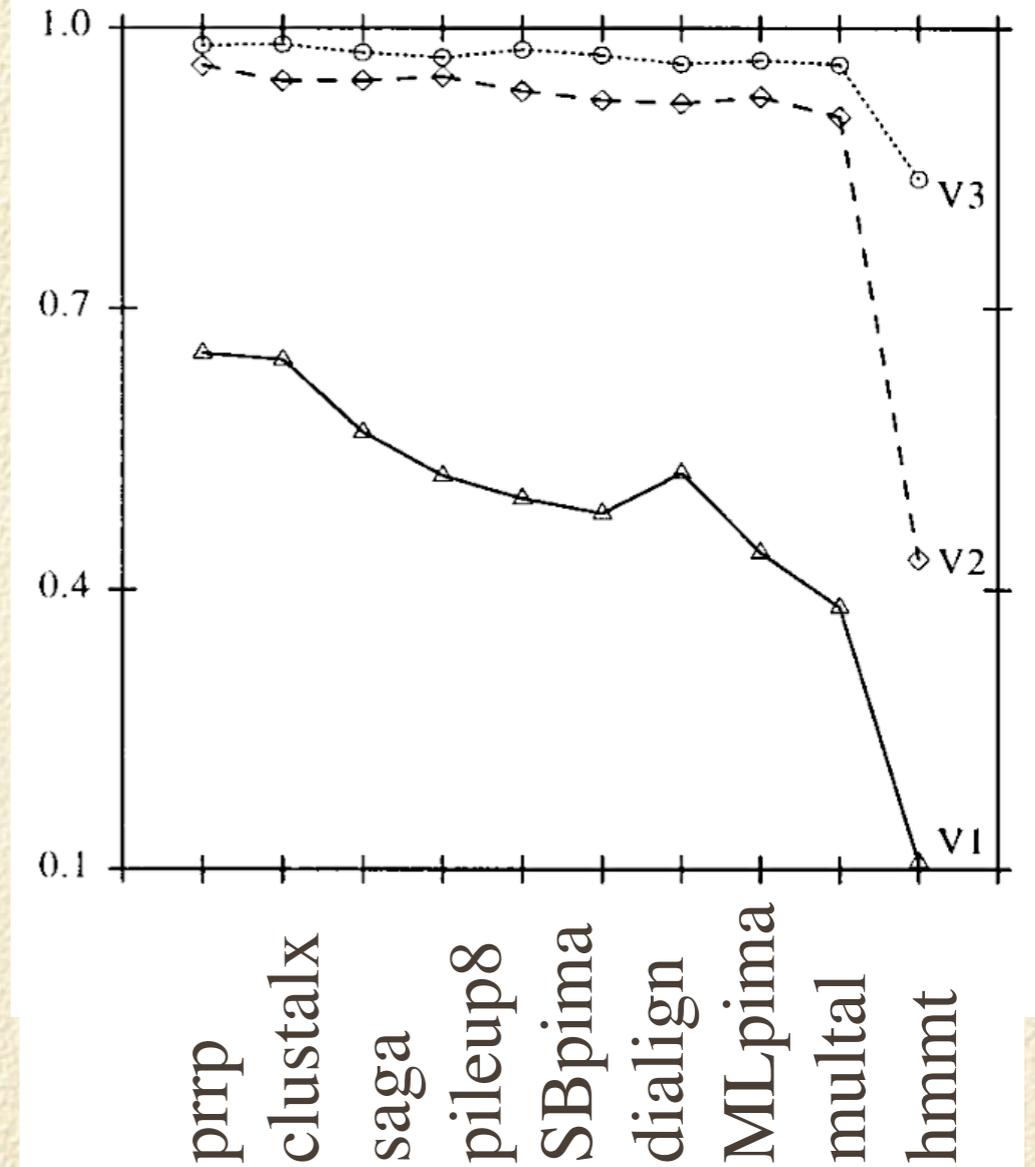
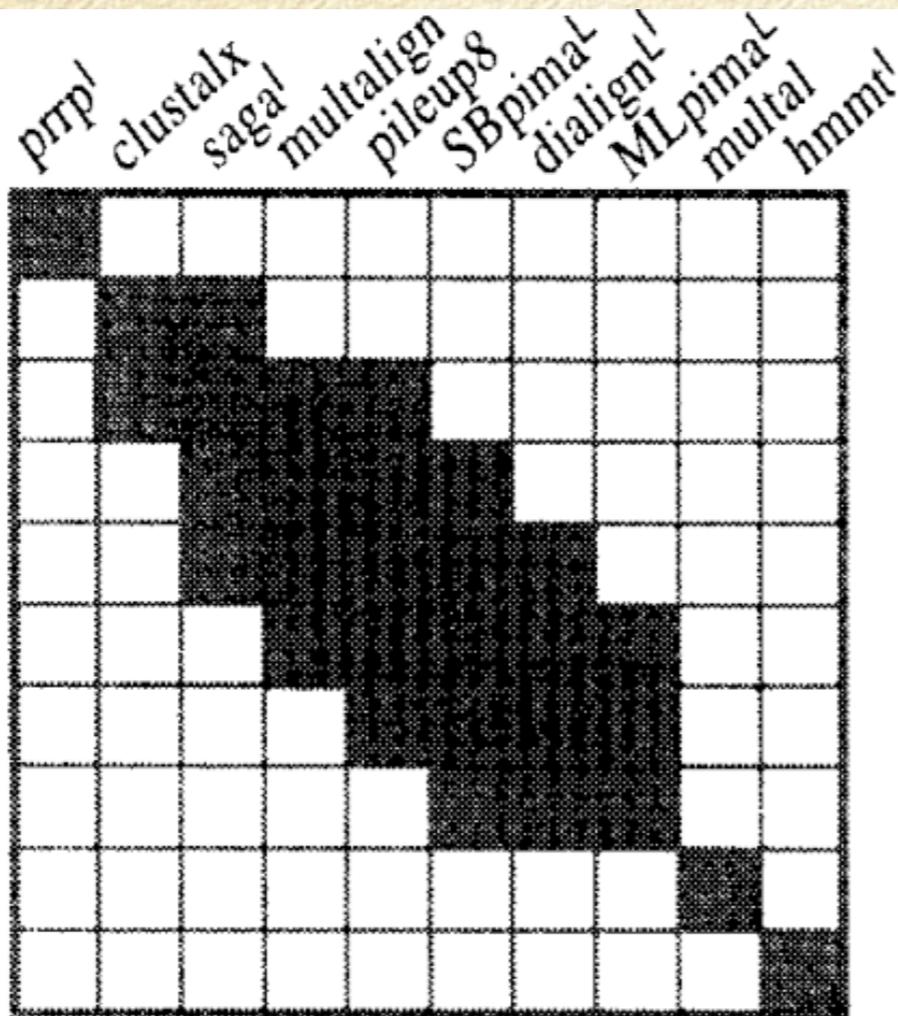
*142 reference sequences*

---

*[http://www-igbmc.u-strasbg.fr/BioInfo/BALiBASE/prog\\_scores.html](http://www-igbmc.u-strasbg.fr/BioInfo/BALiBASE/prog_scores.html)*

	<b>Residues</b>	<b>&lt;100</b>	<b>200&lt;300</b>	<b>&gt;500</b>
<b>Reference 1</b>	<b>&lt;25% identity</b>	<b>7</b>	<b>8</b>	<b>8</b>
	<b>20-40% identity</b>	<b>10</b>	<b>9</b>	<b>10</b>
	<b>&gt;35% identity</b>	<b>10</b>	<b>10</b>	<b>8</b>
<b>Reference 2</b>	<b>homogenous + outlier</b>	<b>9</b>	<b>8</b>	<b>7</b>
<b>Reference 3</b>	<b>2 distantly related sets</b>	<b>5</b>	<b>3</b>	<b>5</b>
		<b>extensions (&lt;400)</b>	<b>inserts (&lt;100)</b>	
<b>Reference 4/5</b>		<b>12</b>	<b>12</b>	

Program	Rank Sum
prrp	234.0
clustalx	316.5
saga	371.5
pileup8	398.0
SBpima	416.0
dialign	448.5
MLpima	465.5
multal	477.0
hmmt	544.0
	785.0



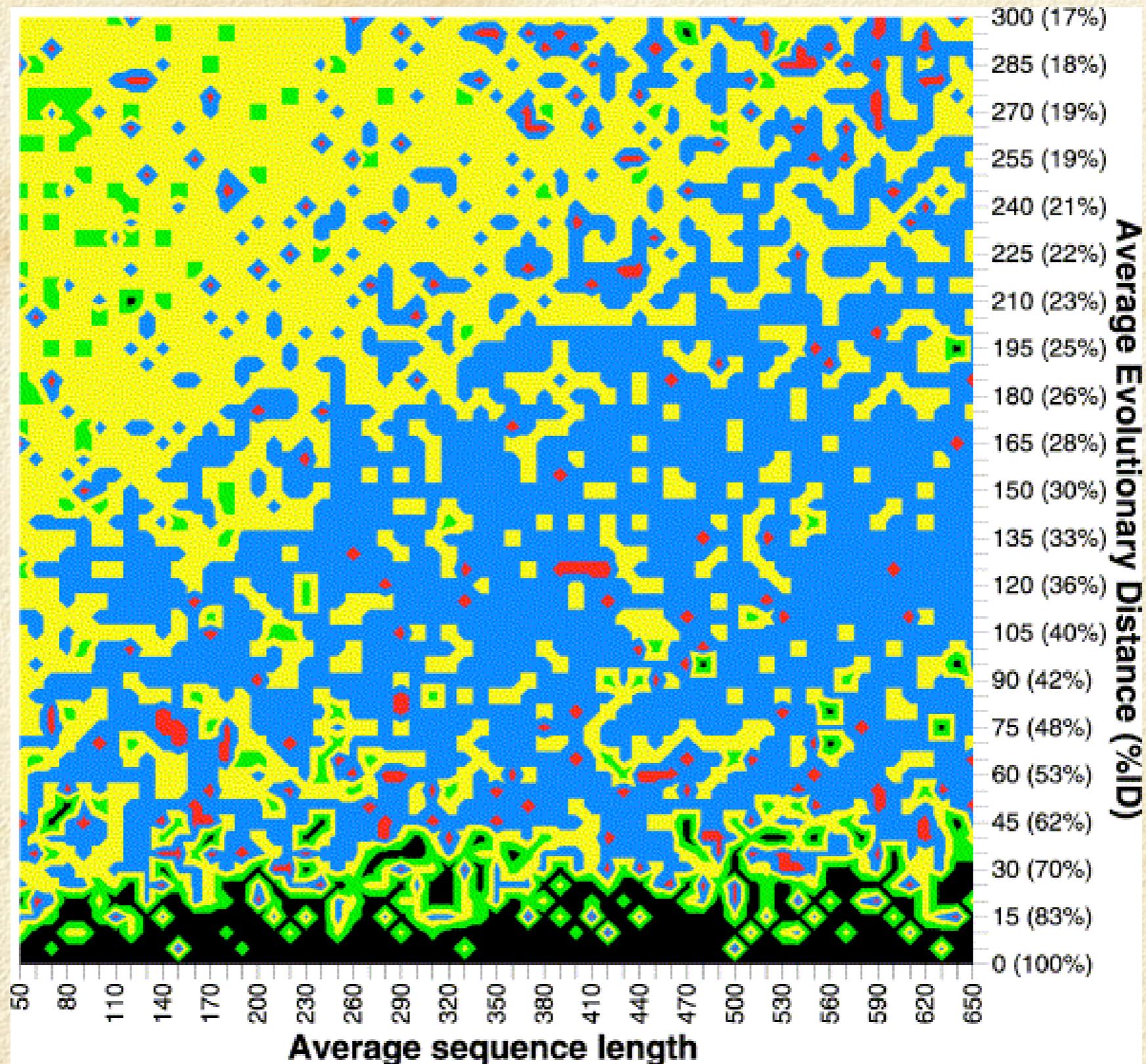


Fig. 1. Color coded matrix showing which method performed best for each pair-combination of conditions: average sequence length (x-axis) and average evolutionary distance (y-axis). The methods are **Poa** (green), **Dialign** (yellow), **T-Coffee** (blue) and **ClustalW** (red).

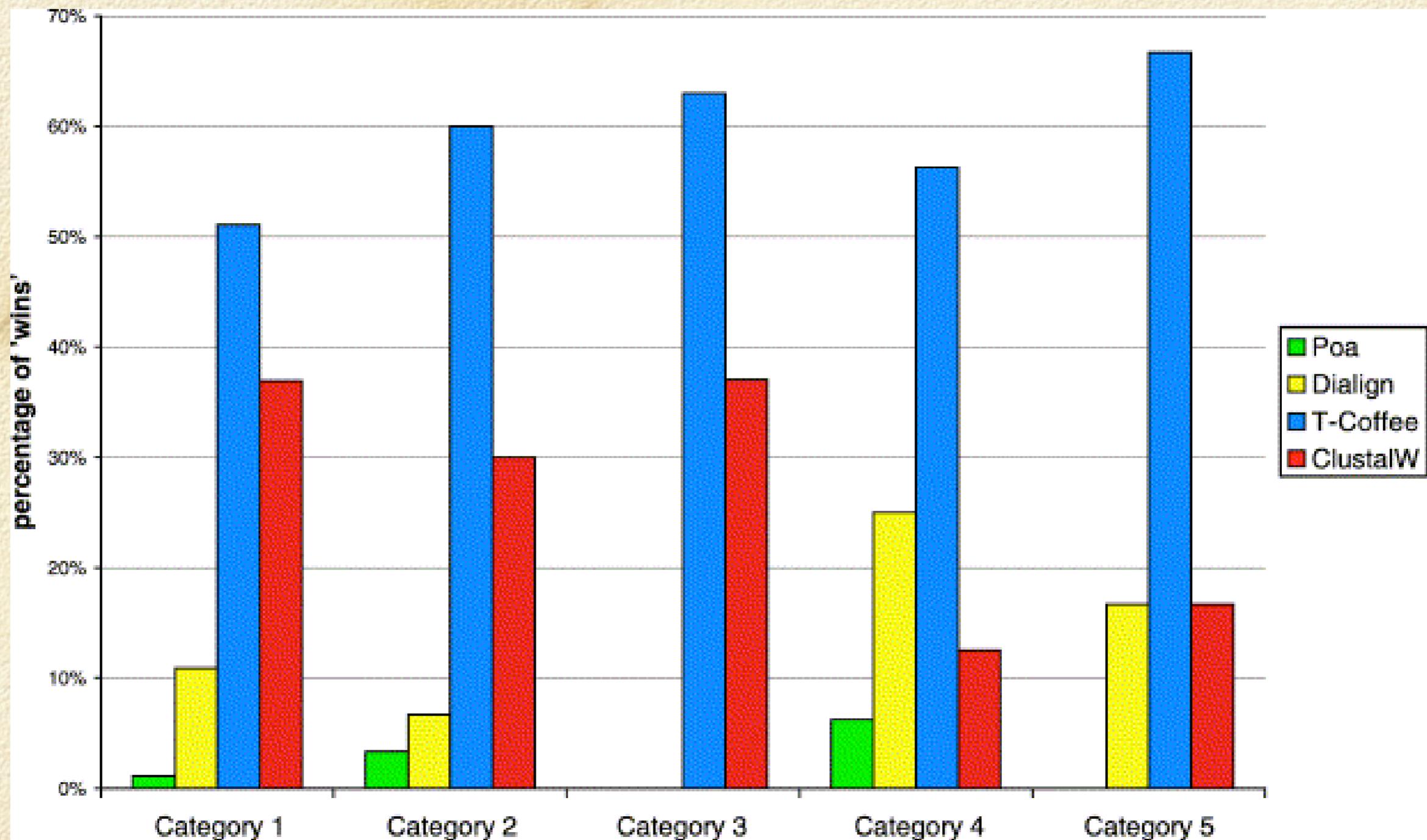
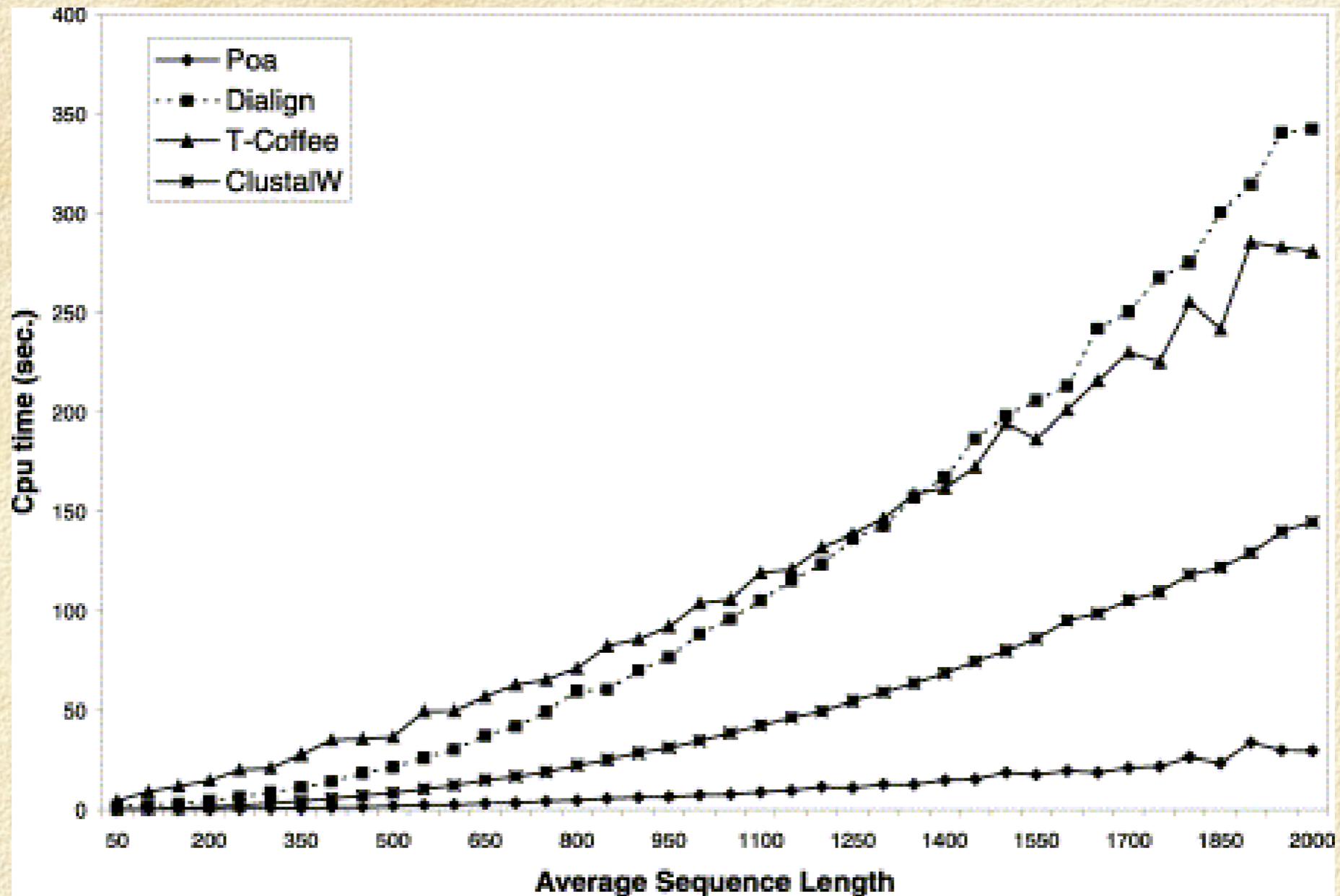


Fig. 2. Results of BALiBASE testing, showing the fraction that each program had the best accuracy (SPS) in each of the five BALiBASE categories.



CPU time consumed by each program to align sets of increasingly long sequences.

# The Problem

153 protein (220AA in length)

- ❑ ClustalW - six minutes
- ❑ T-Coffee - two days
- ❑ MSA - impractical

# Recommendations

---

- **MSA** - for few, short sequences
- **ClustalW** - more versatile, most widely used and only program that *can use multiprocessors*
- **DiAlign** may do better for some
- **T-Coffee** sometimes better than ClustalW, but more computationally expensive
- **POA**, and **MAFFT** new programs which promise speed.

```

dha2_yeast_  -----VSEKSDHDDDKAVVDI SERGRLLNLLADLIERDRDILAAIEHLDNGKPFDEAY
dhac_mouse_  VDKAVKAARQAFQIGSPWRTMDASERGC LLNKLADLMERDRLLLATMEALNGGKVFANAY
dha5_yeast_  VDKAVKAARAAF--DNVWSKTSSEQRGIYLSNLLKLIIEEQDTLAALETLDAGKPYHSNA
dhal_ecoli_  IDRAMSAARGVFE-RGDWSLSSPAKRKAVLNKLADLMEAHAEELALLETLDTGKPIRHSL

dha2_yeast_  LLDLASVLKELRYTAGWADKLGHTLRFAITIPTFQDLRFLRYTRHEPVGVCGEIIPWNIP
dhac_mouse_  LSDLGGCIKALKYCAGWADKIHG-----QTIPSDGDI-FT-YTRREPIGVCGQIIPWNFP
dha5_yeast_  KGDLAQILQLTRYFAGSADKFDKG----ATIPLTFNK-FA-YTLKVPFGVVAQIVPWNYP
dhal_ecoli_  RDDIPGAARAIRWYAEAIDKVYG-----EVATTSSHELA-MIVREPVGVIAAIVPWNFP

dha2_yeast_  LLMYIWKIGPALAAGNTVVLKPEELTPLTALT VATLIKEAGFPPGVVNVVSGYGPTAGAA
dhac_mouse_  MLMFIWKIGPALSCGNTVVVKPAEQTPLTALH LASLIKEAGFPPGVVNIIVPGYGPTAGAA
dha5_yeast_  LAMACWKLQGALAAGNTVI IKPAENTSLSLLYFATLIKAGFPPGVVNIIVPGYGLVGOA
dhal_ecoli_  LLLTCWKLGPALAAGNSVILKPSEKSPLSAIRLAGLAKEAGLPDGVLNVTGFGHEAGQA

dha2_yeast_  CLSHKDNNDKLAFTGSTLVGKVVMAA AAKSNLKKVTLELGGKSPMIVFIDA-DLDWAVENA
dhac_mouse_  ISSHMDVDKVAFTGSTQVGKLIKEAAGKSNLKRVTLELGGKSPCIVFADA-DLDIAVEFA
dha5_yeast_  LASHMDIDKISFTGSTKVGGFVLEASGQSNLKDVTLECGGKSPALVFEDA-DLDKAIDWI
dhal_ecoli_  LSRHNDIDAIAFTGSTRTGKQLLKDAGDSNMKRVWLEAGGKSANIVFADCPDLQQAASAT

dha2_yeast_  HFGVFFNQGCCIAQSRITVHESIYDEIVERDLEKAKKQ--VLG--NPFESDTRYGPQIL
dhac_mouse_  HHGVFYHQGCCVAASRIFVEESVYDEFVKRSVERAKKY--VLG--NPLTPGINQGPQID
dha5_yeast_  AAGIFYNSGQNCNTANSRVYVQSSYDKFVEKFKETAKKEWDVAGKFDPFDEKCI VGPVIS
dhal_ecoli_  AAGIFYNQGVCIAGTRLLEESIADEF LALLKQQAQNW--QPG--HPLDPATTMGTLID

dha2_yeast_  KIEFDSIPRLINSAKAEG--AKVLCGGGRDDSCVGYI IQPTVFADVTDEMRIAKEEIFGP
dhac_mouse_  KEQHDKILD LIESGKKEG--AKLECGGGRWGN-KGFFVQPTVFSNVTDEMRIAKEEIFGP
dha5_yeast_  STQYDRIKSYIERGKREKLD MFQTS EFPIGGAKGYFIPPTIFTDVPQTSKLLQDEIFGP
dhal_ecoli_  CAHADSVHSFIREGESKG----QLLLDGRNAG-LAAAI GPTIFVDVDPNASLSREEIFGP

dha2_yeast_  VITISRFKSVDEAIKRVDNTKYGLAAYVFTK--DKAIRISAALKAGTVVWNCVHVASYQI
dhac_mouse_  VQQIMKFKSVDDVIKRANNTTYGLAAGLFTKDLDKAITVSSALQAGVWVWNCYIMLSAQC
dha5_yeast_  VVVVSKFTNYDDALKLANDTCYGLASAVFTKDVKKAHMFARDIKAGTVWINSNDEDVTV
dhal_ecoli_  VLVVTRFTSEEQALQLANDSQYGLGAAVWTRDLSRAHRMSRRLKAGSVFVNNYNDGDMTV

dha2_yeast_  PFGGNKNSGMGRELGEYGLE-----
dhac_mouse_  PFGGFKMSGNGRELGEHGLYEYTELKTVAMKISQKNS
dha5_yeast_  PFGGFKMSGIGRELGQSGVDTYLQTKAVHINLSLDN-
dhal_ecoli_  PFGGYKQSGNGRDKSLHALEKFTTELKTIWI-----

```

```

dha2_yeast_  -----VSEKSOHDDDKAVVDISERGRLLNILLADLIERDRDILAAIEHLDNGKPFDEAY
dhac_mouse_ VDKAVKAARQAFQIGSPWRTMDASERGCLLNKLADLMERDRLLLATMEALNGGKVFANAY
dha5_yeast_ VDKAVKAARAAFND--VWSKTSSEQRGIYLSNLLKLIIEEQDTLAALETLDAGKPYHSNA
dha1_ecoli_  IDRAMSAARGVFERG-DWSLSSPAKRKAVLNKLADLMEAHAEELALLETLDTGKPIRHSL

dha2_yeast_  LLDLASVLKELRYTAGWADKLGHTLRFAITIPTFQDLRFLRYTRHEPVGVCGEIIPWNIP
dhac_mouse_  LSDLGGCIKALKYCAGWADKIHG-----QTIP--SDGDIFTYTRREPIGVCGQIIPWNFP
dha5_yeast_  KGDLAQILQLTRYFAGSADKFDKG----ATIP--LTFNKFAYTLKVPFGVVAQIVPWNYP
dha1_ecoli_  RDDIPGAARAIRWYAEAIDKVYG-----EVAT--TSSHELAMIVREPVGVIAAIVPWNFP

dha2_yeast_  LLMYIWKIGPALAAGNTVVVKPEELTPLTALTVAATLIKEAGFPPGVVNVVSGYGPTAGAA
dhac_mouse_  MLMFIWKIGPALSCGNTVVVKPAEQTPLTALHLASLIKEAGFPPGVVNIIVPGYGPTAGAA
dha5_yeast_  LAMACWKLQGALAAGNTVVIKPAENTSLSLLYFATLIKKAGFPPGVVNIIVPGYGLVGOA
dha1_ecoli_  LLLTCWKLGPALAAGNSVILKPSEKSPLSAIRLAGLAKEAGLPDGVNLNVVTGFGHEAGQA

dha2_yeast_  CLSHKDNNDKLAFTGSTLVGKVVMAAKSNLKKVTLELGGKSPMIVFIDA-DLDWAVENA
dhac_mouse_  ISSHMDVDKVAFTGSTQVGKLIKEAAGKSNLKRVTLELGGKSPCIVFADA-DLDIAVEFA
dha5_yeast_  LASHMDIDKISFTGSTKVGGFVLEASGQSNLKDVTLECGGKSPALVFEDA-DLDKAIDWI
dha1_ecoli_  LSRHNDIDAI AFTGSTRTGKQLLDAGDSNMKRVWLEAGGKSANIVFADCPDLQQAASAT

dha2_yeast_  HFGVFFNQGCCIAQSRTVHESIYDEIVERDLEKAKKQ--VLG--NPFESDTRYGPQIL
dhac_mouse_  HHGVFYHQGCCVAASRIFVEESVYDEFVKRSVERAKKY--VLG--NPLTPGINQGPQID
dha5_yeast_  AAGIFYNSGQNCNTANSRVYVQSSYDKFVEKFKETAKKEWDVAGKFDPFDEKCIVGPVIS
dha1_ecoli_  AAGIFYNQGVCIAGTRLLEESIADEFLLALKQQAQNW--QPG--HPLDPATTMGTLID

dha2_yeast_  KIEFDSIPRLINSAKAEG--AKVLCGGGRDDSCVGYIYIQPTVFADVTDEMRIAKEEIFGP
dhac_mouse_  KEQHDKILD LIESGKKEG--AKLECGGGR-WGNKGGFFVQPTVFSNVTDEMRIAKEEIFGP
dha5_yeast_  STQYDRIKSYIERGKREEKLD MFQTSSEFPIGGAKGYFIPPTIFTDVPQTSKLLQDEIFGP
dha1_ecoli_  CAHADSVHSFIREGESKG---QLLLDGRN--AGLAAAIGPTIFVDVDPNASLSREEIFGP

dha2_yeast_  VITISRFKSVDEAIKRVDNTKYGLAAYVFTK--DKAIRISAALKAGTVVWNCVHVASYQI
dhac_mouse_  VQQIMKFKSVDDVIKRANNTTYGLAAGLFTKDLDKAITVSSALQAGVWVWNCYIMLSAQC
dha5_yeast_  VVVVSKFTNYDDALKLANDTCYGLASAVFTKDVKKAHMFARDIKAGTVWINSNDEDVTV
dha1_ecoli_  VLVVTRFTSEEQALQLANDSQYGLGAAVWTRDLSRAHRMSRRLKAGSVFVNNYNDGDMTV

dha2_yeast_  PFGGNKNSGMGRELGEYGLE-----
dhac_mouse_  PFGGFKMSGNGRELGEHGLYEYTELKTVAMKISQKNS
dha5_yeast_  PFGGFKMSGIGRELGQSGVDTYLQTKAVHINLSLDN-
dha1_ecoli_  PFGGYKQSGNGRDKSLHALEKFTTELKTIWI-----

```

dha2_yeast	.....VSE	KSQHDDDKAV	VDISERGRLL	NILADLIERD	RDILAAIEHL	dha2_yeast	.....VSE	KSQHDDDKAV	VDISERGRLL	NILADLIERD	RDILAAIEHL
dhac_mouse	VDKAVKAARQ	AFQIGSPWRT	MDASERGCLL	NKLADLMERD	RLLLATMEAL	dhac_mouse	VDKAVKAARQ	AFQIGSPWRT	MDASERGCLL	NKLADLMERD	RLLLATMEAL
dha5_yeast	VDKAVKAARA	AFDN..VWSK	TSSEQRGIYL	SNLLKLIIEE	QDTLAALET	dha5_yeast	VDKAVKAARA	AF..DNVWSK	TSSEQRGIYL	SNLLKLIIEE	QDTLAALET
dha1_ecoli	IDRAMSAARG	VFERG.DWSL	SSPAKRKAVL	NKLADLMEAH	AEELALLETL	dha1_ecoli	IDRAMSAARG	VFE.RGDWSL	SSPAKRKAVL	NKLADLMEAH	AEELALLETL
dha2_yeast	DNGKPFDEAY	LLDLASVLKE	LRYTAGWADK	LHGTLRFAIT	IPTFQDLRFL	dha2_yeast	DNGKPFDEAY	LLDLASVLKE	LRYTAGWADK	LHGTLRFAIT	IPTFQDLRFL
dhac_mouse	NGGKVFANAY	LSDLGGCIKA	LKYCAGWADK	IHG.....QT	IP..SDGDIF	dhac_mouse	NGGKVFANAY	LSDLGGCIKA	LKYCAGWADK	IHG.....QT	IPSDGDI.FT
dha5_yeast	DAGKPYHSNA	KGDLAQILQL	TRYFAGSADK	FDKG....AT	IP..LTFNKF	dha5_yeast	DAGKPYHSNA	KGDLAQILQL	TRYFAGSADK	FDKG....AT	IPLTFNK.FA
dha1_ecoli	DTGKPIRHSL	RDDIPGAARA	IRWYAEAIDK	VYG.....EV	AT..TSSHEL	dha1_ecoli	DTGKPIRHSL	RDDIPGAARA	IRWYAEAIDK	VYG.....E	VATTSSHELA
dha2_yeast	RYTRHEPVG	CGEIIPWNIP	LLMYIWKIGP	ALAAGNTVVL	KPEELTPLTA	dha2_yeast	RYTRHEPVG	CGEIIPWNIP	LLMYIWKIGP	ALAAGNTVVL	KPEELTPLTA
dhac_mouse	TYTRREPIGV	CGQIIPWNFP	MLMFIWKIGP	ALSCGNTVVV	KPAEQTPLTA	dhac_mouse	.YTRREPIGV	CGQIIPWNFP	MLMFIWKIGP	ALSCGNTVVV	KPAEQTPLTA
dha5_yeast	AYTLKVPFGV	VAQIVPWNY	LAMACWKLQG	ALAAGNTVII	KPAENTSLSL	dha5_yeast	.YTLKVPFGV	VAQIVPWNY	LAMACWKLQG	ALAAGNTVII	KPAENTSLSL
dha1_ecoli	AMIVREPVG	IAAIVPWNF	LLLTCWKLGP	ALAAGNSVIL	KPSEKSPLSA	dha1_ecoli	.MIVREPVG	IAAIVPWNF	LLLTCWKLGP	ALAAGNSVIL	KPSEKSPLSA
dha2_yeast	LTVATLIKEA	GFPPGVVNVV	SGYGPTAGAA	CLSHKDNDKL	AFTGSTLVGK	dha2_yeast	LTVATLIKEA	GFPPGVVNVV	SGYGPTAGAA	CLSHKDNDKL	AFTGSTLVGK
dhac_mouse	LHLASLIKEA	GFPPGVVNI	PGYGPTAGAA	ISSHMDVDKV	AFTGSTQVGK	dhac_mouse	LHLASLIKEA	GFPPGVVNI	PGYGPTAGAA	ISSHMDVDKV	AFTGSTQVGK
dha5_yeast	LYFATLIKKA	GFPPGVVNI	PGYGLVQQA	LASHMDIDKI	SFTGSTKVGG	dha5_yeast	LYFATLIKKA	GFPPGVVNI	PGYGLVQQA	LASHMDIDKI	SFTGSTKVGG
dha1_ecoli	IRLAGLAKEA	GLPDGVLNVV	TGFGHEAGQA	LSRHNDIDAI	AFTGSTRTGK	dha1_ecoli	IRLAGLAKEA	GLPDGVLNVV	TGFGHEAGQA	LSRHNDIDAI	AFTGSTRTGK
dha2_yeast	VVMKAAAKSN	LKKVTLELGG	KSPMIVFIDA	.DLDWAVENA	HFGVFFNQGG	dha2_yeast	VVMKAAAKSN	LKKVTLELGG	KSPMIVFIDA	.DLDWAVENA	HFGVFFNQGG
dhac_mouse	LIKEAAGKSN	LKRVTLELGG	KSPCIVFADA	.DLDIAVEFA	HHGVFVHQGG	dhac_mouse	LIKEAAGKSN	LKRVTLELGG	KSPCIVFADA	.DLDIAVEFA	HHGVFVHQGG
dha5_yeast	FVLEASGQSN	LKDVTLECGG	KSPALVFEDA	.DLDKAIDWI	AAGIFYNSGQ	dha5_yeast	FVLEASGQSN	LKDVTLECGG	KSPALVFEDA	.DLDKAIDWI	AAGIFYNSGQ
dha1_ecoli	QLLKDAGDSN	MKRWLEAGG	KSANIVFADC	PDLQQAASAT	AAGIFYNQGG	dha1_ecoli	QLLKDAGDSN	MKRWLEAGG	KSANIVFADC	PDLQQAASAT	AAGIFYNQGG
dha2_yeast	CCIAQSRTIV	HESIYDEIVE	RDLEKAKKQ.	.VLG..NPFE	SDTRYGPQIL	dha2_yeast	CCIAQSRTIV	HESIYDEIVE	RDLEKAKKQ.	.VLG..NPFE	SDTRYGPQIL
dhac_mouse	CCVAASRIFV	EESVYDEFVK	RSVERAKKY.	.VLG..NPLT	PGINQGPQID	dhac_mouse	CCVAASRIFV	EESVYDEFVK	RSVERAKKY.	.VLG..NPLT	PGINQGPQID
dha5_yeast	NCTANSRVVY	QSSIIDKFVE	KFKETAKKEW	DVAGKFDPDF	EKCIVGPVIS	dha5_yeast	NCTANSRVVY	QSSIIDKFVE	KFKETAKKEW	DVAGKFDPDF	EKCIVGPVIS
dha1_ecoli	VCIAGTRLLL	EESIADEFLA	LLKQQAQNW.	.QPG..HPLD	PATTMGTLID	dha1_ecoli	VCIAGTRLLL	EESIADEFLA	LLKQQAQNW.	.QPG..HPLD	PATTMGTLID
dha2_yeast	KIEFDSIPRL	INSAKAEG..	AKVLCGGGRD	DSCVGYIIP	TVFADVDEM	dha2_yeast	KIEFDSIPRL	INSAKAEG..	AKVLCGGGRD	DSCVGYIIP	TVFADVDEM
dhac_mouse	KEQHDKILD	IESGKKEG..	AKLECGGGR.	WGNKGFVQP	TVFSNVDEM	dhac_mouse	KEQHDKILD	IESGKKEG..	AKLECGGGRW	GN.KGFVQP	TVFSNVDEM
dha5_yeast	STQYDRIKSY	IERGKREEKL	DMFQTSEFPI	GGAKGYFIPP	TIFDVPQTS	dha5_yeast	STQYDRIKSY	IERGKREEKL	DMFQTSEFPI	GGAKGYFIPP	TIFDVPQTS
dha1_ecoli	CAHADSVHSF	IREGESKG..	.QLLLDGRN.	.AGLAAIIP	TIFVDVDPNA	dha1_ecoli	CAHADSVHSF	IREGESKG..	.QLLLDGRN	AG.LAAIIP	TIFVDVDPNA
dha2_yeast	RIAKEEIFGP	VITISRFKSV	DEAIKRVNT	KYGLAAYVFT	K..DKAIRIS	dha2_yeast	RIAKEEIFGP	VITISRFKSV	DEAIKRVNT	KYGLAAYVFT	K..DKAIRIS
dhac_mouse	RIAKEEIFGP	VQQIMKFKSV	DDVIKRNNT	TYGLAAGLFT	KDLDKAITVS	dhac_mouse	RIAKEEIFGP	VQQIMKFKSV	DDVIKRNNT	TYGLAAGLFT	KDLDKAITVS
dha5_yeast	KLLQDEIFGP	VVVVSKFTNY	DDALKLANDT	CYGLASAVFT	KDVKKAHMFA	dha5_yeast	KLLQDEIFGP	VVVVSKFTNY	DDALKLANDT	CYGLASAVFT	KDVKKAHMFA
dha1_ecoli	SLSREEIFGP	VLVWTRFTSE	EQALQLANDS	QYGLGAAWWT	RDLRAHRMS	dha1_ecoli	SLSREEIFGP	VLVWTRFTSE	EQALQLANDS	QYGLGAAWWT	RDLRAHRMS
dha2_yeast	AALKAGTWWV	NCVHVASYQI	PFGGNKNSGM	GRELGEYGLE	.....	dha2_yeast	AALKAGTWWV	NCVHVASYQI	PFGGNKNSGM	GRELGEYGLE	.....
dhac_mouse	SALQAGVWWV	NCYIMLSAQC	PFGGFKMSGN	GRELGEHGLY	EYTELKTIVAM	dhac_mouse	SALQAGVWWV	NCYIMLSAQC	PFGGFKMSGN	GRELGEHGLY	EYTELKTIVAM
dha5_yeast	RDIKAGTWWI	NSSNDEDVTV	PFGGFKMSGI	GRELQSGVD	TYLQTKAVHI	dha5_yeast	RDIKAGTWWI	NSSNDEDVTV	PFGGFKMSGI	GRELQSGVD	TYLQTKAVHI
dha1_ecoli	RRLKAGSVFV	NNYNDGDMTV	PFGGYKQSGN	GRDKSLHALE	KFTELKTIWI	dha1_ecoli	RRLKAGSVFV	NNYNDGDMTV	PFGGYKQSGN	GRDKSLHALE	KFTELKTIWI

dha2_yeast_	-----VSEKSDHDDDKAVVDISERGRLLNILADLIERDRDILAAIEHLDNGKPFDEAY
dhac_mouse_	VDKAVKAARQAFQIGSPWRTMDASERGCLLNKLADLMERDRLLLATMEALNGGKVFANAY
dha5_yeast_	VDKAVKAARAAF--DNVWSKTSSEQRGIYLSNLLKLIIEEQDTLAALETLDAGKPYHSNA
dhal_ecoli_	IDRAMSAARGVFE-RGDWSLSSPAKRKAVLNKLADLMEAHAEEELALLETLDTGKPIRHSL
dha2_yeast_	LLDLASVLKELRYTAGWADKLHGTLRFAITIPTFQDLRFLRYTRHEPVGVCGEIIPWNIP
dhac_mouse_	LSDLGGCIKALKYCAGWADKIHG-----QTIPSDGDI-FT-YTRREPIGVCGQIIPWNFP
dha5_yeast_	KGDLAQILQLTRYFAGSADKFDKG-----ATIPLTFNK-FA-YTLKVPFGVVAQIVPWNYP
dhal_ecoli_	RDDIPGAARAIRWYAEAIDKVYG-----EVATTSSELA-MIVREPVGVIAAIVPWNFP
dha2_yeast_	LLMYIWKIGPALAAGNTVVLKPEELTPLTALTVAATLIKEAGFPPGVVNVVSGYGPTAGAA
dhac_mouse_	MLMFIWKIGPALSCGNTVVVKPAEQTPLTALHLASLIKEAGFPPGVVNIVPGYGPTAGAA
dha5_yeast_	LAMACWKLQGALAAGNTVVIKPAENTSLSLLYFATLIKKAGFPPGVVNIVPGYGSLVGQA
dhal_ecoli_	LLLTCWKLGPALAAGNSVILKPSEKSPLSAIRLAGLAKEAGLPDGVLNVVTGFGHEAGQA
dha2_yeast_	CLSHKDNDKLAFTGSTLVGKVVMKAAAKSNLKKVTLELGGKSPMIVFIDA-DLDWAVENA
dhac_mouse_	ISSHMDVDKVAFTGSTQVGKLIKEAAGKSNLKRVTLELGGKSPCIVFADA-DLDIAVEFA
dha5_yeast_	LASHMDIDKISFTGSTKVGGFVLEASGQSNLKDVTLECGGKSPALVFEDA-DLDKAIDWI
dhal_ecoli_	LSRHNDIDAI AFTGSTRTGKQLLDAGDSNMKRVWLEAGGKSANIVFADCPDLQQAASAT
dha2_yeast_	HFGVFFNQGCCIAQSRITVHESIYDEIVERDLEKAKKQ--VLG--NPFESDTRYGPQIL
dhac_mouse_	HHGVFYHQGCCVAASRIFVEESVYDEFVKRSVERAKKY--VLG--NPLTPGINQGPQID
dha5_yeast_	AAGIFYNSGQNCTANSRVYVQSSIYDKFVEKFKETAKKEWDVAGKFDPFDEKCI VGPVIS
dhal_ecoli_	AAGIFYNQGVCIAGTRLLLEESIADEFLALLKQQAQNW--QPG--HPLDPATTMGTLID
dha2_yeast_	KIEFDSIPRLINSAKAEG--AKVLCGGGRDDSCVGYIYIQPTVFADVTDDEMRIAKEEIFGP
dhac_mouse_	KEQHDKILDIESGKKEG--AKLECGGGRWGN-KGFFVQPTVFSNVTDEMRIAKEEIFGP
dha5_yeast_	STQYDRIKSYIERGKREEKLDMFQTSEFPIGGAKGYFIPPTIFTDVPQTSKLLQDEIFGP
dhal_ecoli_	CAHADSVHSFIREGESKG---QLLLDGRNAG-LAAAIGPTIFVDVDPNASLSREEIFGP
dha2_yeast_	VITISRFKSVDEAIKRVDNTKYGLAAYVFTK--DKAIRISAALKAGTVWVNCVHVASYQI
dhac_mouse_	VQQIMKFKSVDDEVIKRANNTTYGLAAGLFTKDLDKAITVSSALQAGVWVNCYIMLSAQC
dha5_yeast_	VVVVSKFTNYDDALKLANDTCYGLASAVFTKDVKKAHMFARDIKAGTVWINSNDEDVTV
dhal_ecoli_	VLVVTRFTSEEQALQLANDS QYGLGAAVWTRDL SRAHRMSRRLKAGSVFVNNDGDMTV
dha2_yeast_	PFGGNKNSGMGRELGEYGLE-----
dhac_mouse_	PFGGFKMSGNGRELGEHGLYEYTELKT VAMKISQKNS
dha5_yeast_	PFGGFKMSGIGRELQSGVD TYLQTKAVHINLSLDN-
dhal_ecoli_	PFGGYKQSGNGRDKSLHALEKFTTELKTIWI-----

```

dha2_yeast_  -----VSEKSQHDDDKAVVDI SERGRLLNI LADLIERDRDI LAAIEHLDNGKPFDEAY
dhac_mouse_ VDKAVKAARQAFQIGSPWRTMDASERGC LLNKLADLMERDRLL LATMEALNGGKVFANAY
dha5_yeast_ VDKAVKAARAAF--DNVWSKTSSEQRGIYLSNLLKLI EEEQDTLAALETLDAGKPYHSNA
dha1_ecoli_  IDRAMSAARGVFE-RGDWSLSSPAKRKAVLNKLADLMEAHAEELALLETLDTGKPIRHSL

dha2_yeast_  LLDLASVLKELRYTAGWADKLGHTLRFAI TIPTFQDLRFLRYTRHEPVGVCGEIIPWNIIP
dhac_mouse_  LSDLGGCIKALKYCAGWADKIHG-----QTIPSDGDI-FT-YTRREPIGVCGQIIPWNFP
dha5_yeast_  KGDLAQILQLTRYFAGSADKFDKG----ATIPLTFNK-FA-YTLKVPFGVVAQIVPWNYP
dha1_ecoli_  RDDIPGAARAIRWYAEAIKVVYG-----EVATTSSHELA-MIVREPVGVIAAIVPWNFP

dha2_yeast_  LLMYIWKIGPALAAGNTVVLKPEELTPLTALT VATLIKEAGFPPGVVNVVSGYGPTAGAA
dhac_mouse_  MLMFIWKIGPALSCGNTVVVKPAEQTPLTALHLASLIKEAGFPPGVVNIIVPGYGPTAGAA
dha5_yeast_  LAMACWKLQGALAAGNTVI IKPAENTSLSLLYFATLIKKAGFPPGVVNIIVPGYGLVGOA
dha1_ecoli_  LLLTCWKLGPALAAGNSVILKPSEKSPLSAIRLAGLAKEAGLPDGVLVNVVTGFGHEAGQA

dha2_yeast_  CLSHKDNDKLAFTGSTLVGKVVMMKAAAKSNLKKVTLELGGKSPMIVFIDA-DLDWAVENA
dhac_mouse_  ISSHMDVDKVAFTGSTQVGKLIKEAAGKSNLKRVTLELGGKSPCIVFADA-DLDIAVEFA
dha5_yeast_  LASHMDIDKISFTGSTKVGGFVLEASGQSNLKDVTLECGGKSPALVFEDA-DLDKAIDWI
dha1_ecoli_  LSRHNDIDAI AFTGSTRTGKQLLKDAGDSNMKRVWLEAGGKSANIVFADCPDLQQAASAT

dha2_yeast_  HFGVFFNQGQCCIAQSRITVHESIYDEIVERDLEKAKKQ--VLG--NPFESDTRYGPQIL
dhac_mouse_  HHGVFYHQGCCVAASRIFVEESVYDEFVKRSVERAKKY--VLG--NPLTPGINQGPQID
dha5_yeast_  AAGIFYNSGQNCTANSRVYVQSSIYDKFVEKFKETAKKEWDVAGKFDPFDEKCIVGPVIS
dha1_ecoli_  AAGIFYNQGVCIAGTRLLLEESIADEFLALLKQQAQNW--QPG--HPLDPATTMGTLID

dha2_yeast_  KIEFDSIPRLINSAKAEG--AKVLCGGGRDDSCVGYIYIQPTVFADVTDDEMRIAKEEIFGP
dhac_mouse_  KEQHDKILD LIESGKKEG--AKLECGGGRWGN-KGFFVQPTVFSNVTDEMRIAKEEIFGP
dha5_yeast_  STQYDRIKSYIERGKREEKLD MFQTS EFPIGGAKGYFIPPTIFTDVPQTSKLLQDEIFGP
dha1_ecoli_  CAHADSVHSFIREGESKG----QLLLDGRNAG-LAAAIGPTIFVDVDPNASLSREEIFGP

dha2_yeast_  VITISRFKSVDEAIKRVDNTKYGLAAYVFTK--DKAIRISAALKAGTVVWNCVHVASYQI
dhac_mouse_  VQQIMKFKSVDDVIK RANNTTYGLAAGLFTKDLDKAITVSSALQAGVVWVNCYIMLSAQC
dha5_yeast_  VVVVSKFTNYDDALKLANDTCYGLASAVFTKDVKKAHMFARDIKAGTVWINSNDEDVTV
dha1_ecoli_  VLVVTRFTSEEQALQLANDSQYGLGAAVWTRDLSRAHRMSRRLKAGSVFVNNYNDGDMTV

dha2_yeast_  PFGGNKNSGMGRELGEYGLE-----
dhac_mouse_  PFGGFKMSGNGRELGEHGLYEYTELKTVAMKISQKNS
dha5_yeast_  PFGGFKMSGIGRELGQSGVDTYLQTKAVHINLSLDN-
dha1_ecoli_  PFGGYKQSGNGRDKSLHALEKFTTELKTIWI-----

```



dha2_yeast__	-- ----VSEKSOHDDDKAVVDISERGRLLNILADLIERDRDILAAIEHLDNGKPFDEAY
dhac_mouse__	VDKAVKAARQAFQIGSPWRTMDASERGCLLNKLADLMERDRLLLATMEALNGGKVFANAY
dha5_yeast__	VDKAVKAARAAF--DNVWSKTSSEQRGIYLSNLLKLIIEEQDTLAALETLDAGKPYHSNA
dhal_ecoli__	IDRAMSAARGVFE--RGDWSLSSPAKRKAVLNKLADLMEAHAEELALLETLDTGKPIRHSL
dha2_yeast__	LLDLASVLKELRYTAGWADKLGHTLRFAITIPTFQDLRFLRYTRHEPVGVCGEIIPWNIP
dhac_mouse__	LSDLGGCIKALKYCAGWADKIHG-----QTIPSDGDI-FT-YTRREPIGVCGQIIPWNFP
dha5_yeast__	KGDLAQILQLTRYFAGSADKFDKG----ATIPLTFNK-FA-YTLKVPFGVVAQIVPWNYP
dhal_ecoli__	RDDIPGAARAIRWYAEAIDKVYG-----EVATTSSHELA-MIVREPVGVIAAIVPWNFP
dha2_yeast__	LLMYIWKIGPALAAGNTVVLKPEELTPLTALT VATLIKEAGFPPGVVNVVSGYGPTAGAA
dhac_mouse__	MLMFIWKIGPALSCGNTVVVKPAEQTPLTALHLASLIKEAGFPPGVVNIIVPGYGPTAGAA
dha5_yeast__	LAMACWKLQGALAAGNTVVIKPAENTSLSLLYFATLIKKAGFPPGVVNIIVPGYGSLVGQA
dhal_ecoli__	LLLTCWKLGPALAAGNSVILKPSEKSPLSAIRLAGLAKEAGLPDGVLVNVVTGFGHEAGQA
dha2_yeast__	CLSHKDNDKLAFTGSTLVGKVVMMKAAAKSNLKKVTLELGGKSPMIVFIDA-DLDWAVENA
dhac_mouse__	ISSHMDVDKVAFTGSTQVGKLIKEAAGKSNLKRVTLELGGKSPCIVFADA-DLDIAVEFA
dha5_yeast__	LASHMDIDKISFTGSTKVGGFVLEASGQSNLKDVTLECGGKSPALVFEDA-DLDKAIDWI
dhal_ecoli__	LSRHNDIDAI AFTGSTRTGKQLLDAGDSNMKRVWLEAGGKSANIVFADCPDLQQAASAT
dha2_yeast__	HFGVFFNQGCCIAQSRITVHESIYDEIVERDLEKAKKQ--VLG--NPFESDTRYGPQIL
dhac_mouse__	HHGVFYHQGCCVAASRIFVEESVYDEFVKRSVERAKKY--VLG--NPLTPGINQGPQID
dha5_yeast__	AAGIFYNSGQNCTANSRVYVQSSIYDKFVEKFKETAKKEWDVAGKFDPFDEKCIIVGPVIS
dhal_ecoli__	AAGIFYNQGVCIAGTRLLLEESIADEFLALLKQQAQNW--QPG--HPLDPATTMGTLID
dha2_yeast__	KIEFDSIPRLINSAKAEG--AKVLCGGGRDSDCVGYIYIQPTVFADVTDEMRIAKEEIFGP
dhac_mouse__	KEQHDKILD LIESGKKEG--AKLECGGGRWGN-KGFFVQPTVFSNVTDEMRIAKEEIFGP
dha5_yeast__	STQYDRIKSYIERGKREEKLD MFQTS EFPIGGA KGYFIPPTIFTDVPQTSKLLQDEIFGP
dhal_ecoli__	CAHADSVHSFIREGESKG---QLLLDGRNAG-LAAAIGPTIFVDVDPNASLSREEIFGP
dha2_yeast__	VITISRFKSVDEAIKRVDNTKYGLAAYVFTK--DKAIRISAALKAGTVWVNCVHVASYQI
dhac_mouse__	VQQIMKFKSVDDVIKRANNTTYGLAAGLFTKDLDKAITVSSALQAGVWVNCYIMLSAQC
dha5_yeast__	VVVVSKFTNYDDALKLANDTCYGLASAVFTKDVKKAHMFARDIKAGTVWINSNDEDVTV
dhal_ecoli__	VLVVTRFTSEEQALQLANDSQYGLGAAVWTRDLSRAHRMSRRLKAGSVFVNNYNDGDMTV
dha2_yeast__	PFGGNKNSGMGRELGEYGLE-----
dhac_mouse__	PFGGFKMSGNGRELGEHGLYEYTELKT VAMKISQKNS
dha5_yeast__	PFGGFKMSGIGRELQSGVD TYLQTKAVHINLSLDN-
dhal_ecoli__	PFGGYKQSGNGRDKSLHALEKFTTELKTIWI-----

1

60

oth:1bbt3 -----FTNLLDVAEACPTFLRFEGGVPYVTTKTDSDRVLAQFDMSLAAKHMSNTFL  
 oth:1aym3 VKNLIEMCQVDTLIPINSTQSNIGNVSMYTVTLSPQTKLAEEIFAIKVDIASHPLATTLI  
 oth:1bbt2 -----GLETRV---VQAERFFK--THLFDWVTSDSFGRCHLLELPTDH--KGVY  
 oth:1aym1 -----EMSVESFLGRSG--CIHESVLDIVDNYNDQSFTKWNINLQEMAQIRRKFEMFTY  
 oth:1bbt1 -----QHTDVS---FIMDRFVK---VTPQNQ---INILDLMQVPSH---TLV

61

120

oth:1bbt3 AGLAQYYTQYSGTINLHFMFTGPTDAKARYMVAYAPPGMEPPKTPEA----AAHCIHAEW  
 oth:1aym3 GEIASYFTHWTGSLRFSFMFCGTANTTLKVL LAYTPPGIGKPRSRKE----AMLGTHVWV  
 oth:1bbt2 GSLTDSYAYMRNGWDVEVTAVGNQFNNGCLLVAMVPELCSIQKRELY---QLTLFPHQFI  
 oth:1aym1 ARFDSEITMVP-SVAAKDGHIHIGHVMQ----YMYVPPGAPIPTTRDDYAWQSGTNASVFW  
 oth:1bbt1 GGLLRASYYYFSDLEIAVKHEG--D-----LTWVPNGA--PEKALD----NTTNPTAYH

121

180

oth:1bbt3 DTGLNSKFTFSIPYLS--AADYTYTASDVAETTN-VQGWVCL-----FQITHGKAD--  
 oth:1aym3 DVGLQSTVSLVVPWIS--ASQYRFTTPDTYSSAGYITCWYQTN-----FVVPPNTPN--  
 oth:1bbt2 NPRTNMTAHITVPFVG--VNRDQYKVHKPWTLV--VMVVAP-----LTVNTEGAP--  
 oth:1aym1 QHGQPFPR-FSLPFLS IASAYYMFYDGYDGDYKSRVGTVVTNDMGTLCSRIVTSEQLHK  
 oth:1bbt1 KAPLTR---LALPYT--APHRVLATVYNGECXX--XRTLPTS-----FNYGAIKATR-

181

209

oth:1bbt3 ---GDALVVLASAGKDFELRLPVDARAE-  
 oth:1aym3 ---TAEMLCFVSGCKDFCLRMARDTDLHK  
 oth:1bbt2 ---QIKVYANIAPTINVHVAG-EFPS-KE-  
 oth:1aym1 VKVVTRIIYHKAKHTKAWCPR-PPRA-VQ-  
 oth:1bbt1 ---VTELLYRMKRAETYCPR-PLLA-IH-

# ClustalW

```
ATT1_DROME MQKTSILILA---LFAIAEAVP---TTGPIRVRRRQVLGGSLASNPAGGADARLNLSKIGIG
ATTA_DROME MQKTSILIVALVALFAITEALPSLPTTGPPIRVRRRQVLGGSLTNSNPAGGADARLDLTKGIG
SW_P36193 -MKFTIVFLLLACVFAMAVATP-----GKPRP-----YSPRPTSHPRP-IRVRR---

ATT1_DROME NPNHNVVGQVFAAGNTQSGPVTTGGTLAYNNAGHGASLTKTHTPGVKDVFQQEAHANLNFN
ATTA_DROME NPNHNVVGQVFAAGNTQSGPVTTGGTLAYNNAGHGASLTKTHTPGVKDVFQQEAHANLNFN
SW_P36193 --EALAIEDHLAQAAIRPPPILPA-----

ATT1_DROME NGRHNLDKVFASQNKLANGFEFQRNGAGLDYSHINGHGASLTHSNFPGIGQQLGLDGRA
ATTA_DROME NGRHNLDKVFASQNKLANGFEFQRNGAGLDYSHINGHGASLTHSNFPGIGQQLGLDGRA
SW_P36193 -----

ATT1_DROME NLWSSPNRATTDLTGSASKWTSGPFANQKPNFGAGLGLSHHFG
ATTA_DROME NLWSSPNRATTDLTGSASKWTSGPFANQKPNFGAGLGLSHHFG
SW_P36193 -----
```

# T-Coffee

```
ATT1_DROME MQKTSILILAL---FAIAEAVP-----TTG--PIRVRRRQVLGGSLASNPAGGADA
ATTA_DROME MQKTSILIVALVALFAITEALPSL-----PTTG--PIRVRRRQVLGGSLTNSNPAGGADA
SW_P36193 MKFTIVFLLLA-CVFAMAVATPGKPRPYS PRPTSHPRPIRVRREAL-----

ATT1_DROME RLNLSKIGINPNHNVVGQVFAAGNTQSGPVTTGGTLAYNNAGHGASLTKTHTPGVKDVFQ
ATTA_DROME RLDLTKGIGINPNHNVVGQVFAAGNTQSGPVTTGGTLAYNNAGHGASLTKTHTPGVKDVFQ
SW_P36193 -----AIEDHLAQAAIRPPPILPA-----

ATT1_DROME QEAHANLFNNGRHNLDKVFASQNKLANGFEFQRNGAGLDYSHINGHGASLTHSNFPGIG
ATTA_DROME QEAHANLFNNGRHNLDKVFASQNKLANGFEFQRNGAGLDYSHINGHGASLTHSNFPGIG
SW_P36193 -----

ATT1_DROME QQLGLDGRANLWSSPNRATTDLTGSASKWTSGPFANQKPNFGAGLGLSHHFG
ATTA_DROME QQLGLDGRANLWSSPNRATTDLTGSASKWTSGPFANQKPNFGAGLGLSHHFG
SW_P36193 -----
```



# MSA Evaluation

---

- **AltAVisT - A WWW tool for comparison of alternative multiple alignments**

<http://bibiserv.techfak.uni-bielefeld.de/altavist/>

- **T-Coffee Server**

<http://igs-server.cnrs-mrs.fr/Tcoffee/>

- **BaliScore comparison**

<http://genome.nci.nih.gov/tools/msacomp.html>

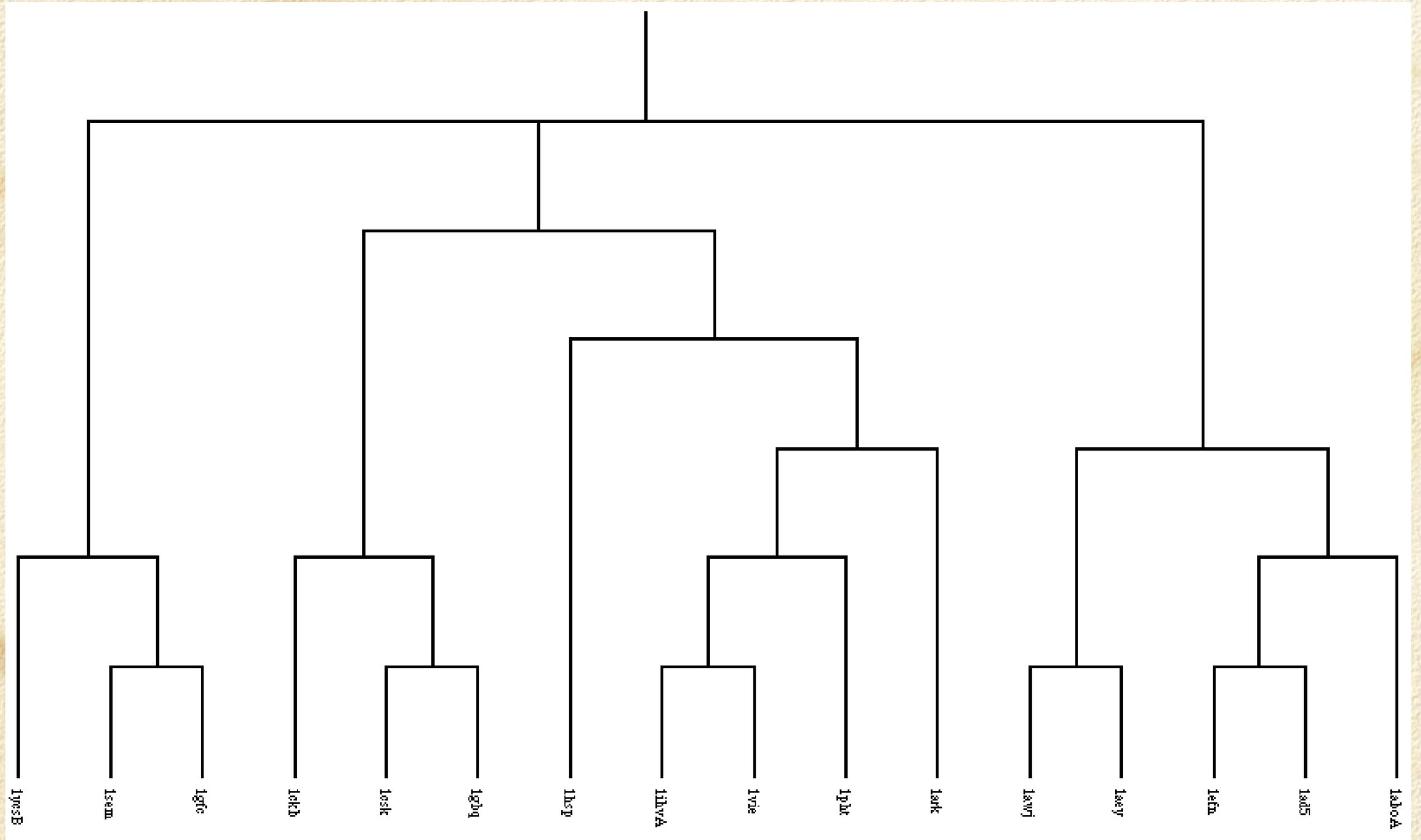
# MSA hurdles

---

- ❑ Too many sequences
- ❑ Repeated sequences are renowned for confusing existing methods
- ❑ MSA methods mostly not parallelized and so still require “super computers”
- ❑ Combine 3D structural info
- ❑ Precomputed families - curated by experts (no need for complete alignment)

# Tree - Dendrogram

*(clustering, not phylogeny)*



# Tree Viewing/Drawing

---

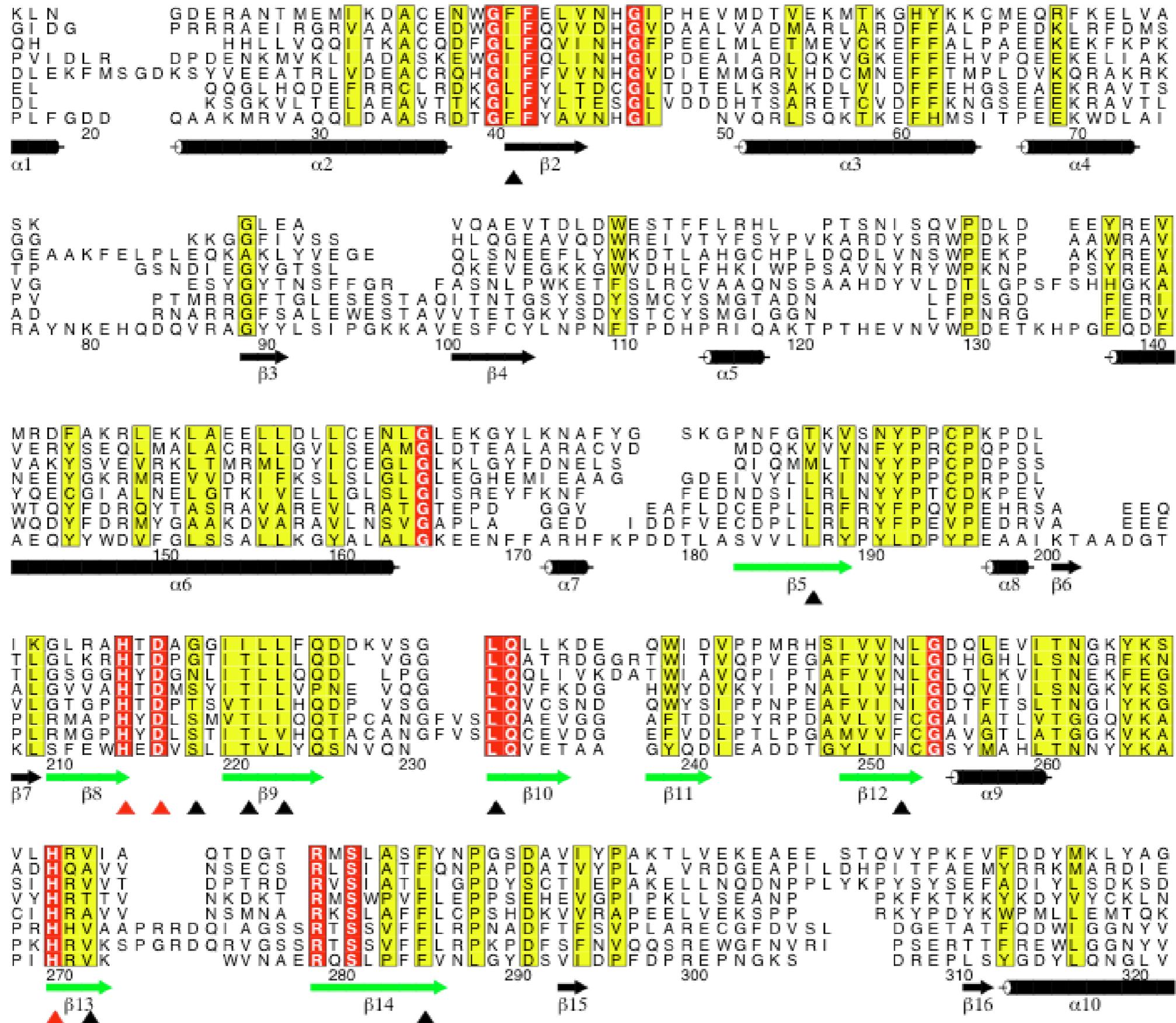
- **PhyloDendron Phylogenetic tree printer**  
<http://iubio.bio.indiana.edu/treeapp/treeprint-form.html>
- **TreeTop** - Phylogenetic Tree Prediction  
[http://www.genebee.msu.su/services/phtree\\_reduced.html](http://www.genebee.msu.su/services/phtree_reduced.html)
- **TreeView** (local view and print)  
<http://taxonomy.zoology.gla.ac.uk/rod/treeview.html>
- **NJPLOT (ClustalW)**  
<ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX>

# Pretty Output

---

- ❑ **Alscript**  
<http://www.compbio.dundee.ac.uk/Software/Alscript/alscript.html>
- ❑ **Pretty EMBOSS**  
<http://www.emboss.org/>
- ❑ **BOXSHADE**  
<http://bioweb.pasteur.fr/seqanal/interfaces/boxshade.html>
- ❑ **ESPrpt**  
<http://prodes.toulouse.inra.fr/ESPrpt/ESPrpt/>
- ❑ **AMAS**  
<http://www.compbio.dundee.ac.uk/amas/>

# Alscript - Output



# Editors

---

- **JalView (J)**  
<http://www.compbio.dundee.ac.uk/Software/JalView/jalview.html>
- **CINEMA (J)**  
<http://bioinf.man.ac.uk/dbbrowser/CINEMA2.1/>
- **Seaview (UMP)**  
<http://pbil.univ-lyon1.fr/software/seaview.html>
- **MPSA (UM)**  
<http://mpsa-pbil.ibcp.fr/>
- **Se-A1 (M)**  
<http://evolve.zoo.ox.ac.uk/software.html?id=seal>
- **ClustalX (UMP)**  
<ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX>

# Multiple Genome Alignment

---

## □ MGA

*Michael Höhl, Stefan Kurtz, Enno Ohlebusch*

*Efficient Multiple Genome Alignment Bioinformatics, Vol. 18 (SI): S312-S320, 2002*

<http://bibiserv.techfak.uni-bielefeld.de/mga/ref.html>

## □ PipMaker and MultiPipMaker

*Schwartz S, Elnitski L, Li M, et al.*

*MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences*

*NUCLEIC ACIDS RES 31 (13): 3518-3524 JUL 1 2003*

<http://bio.cse.psu.edu/pipmaker/>

## □ MAVID

*Bray N and Pachter L, MAVID multiple alignment server, Nucleic Acids Research 2003 31: 3525-3526*

<http://baboon.math.berkeley.edu/mavid/>

<http://www-gsd.lbl.gov/vista/>

# MGA - output

G<sub>1</sub> M G<sub>2</sub> M G<sub>3</sub> MA(MA)x(3)MAM G<sub>4</sub> M G<sub>5</sub> M G<sub>6</sub> M G<sub>7</sub> MAMAM G<sub>8</sub> MA(MA)x(5)MAM G<sub>9</sub> MA(MA)x(87)MAM  
G<sub>10</sub> MA(MA)x(1)MAM G<sub>11</sub> M G<sub>12</sub> MA(MA)x(10)MAM G<sub>13</sub> MA(MA)x(10)MAM G<sub>14</sub> MA(MA)x(12)MAM G<sub>15</sub> MAMAM  
G<sub>16</sub> MAM G<sub>17</sub> MA(MA)x(1)MAM G<sub>18</sub> MA(MA)x(20)MAM G<sub>19</sub> M G<sub>20</sub> MA(MA)x(53)MAM G<sub>21</sub> MA(MA)x(38)MAM G<sub>22</sub>  
M G<sub>23</sub> M G<sub>24</sub> M G<sub>25</sub> M G<sub>26</sub> MA(MA)x(4)MAM G<sub>27</sub> MA(MA)x(1)MAM G<sub>28</sub> M G<sub>29</sub> MA(MA)x(27)MAM G<sub>30</sub>  
MA(MA)x(1)MAM G<sub>31</sub> MA(MA)x(12)MAM G<sub>32</sub> MA(MA)x(3)MAM G<sub>33</sub> MA(MA)x(18)MAM G<sub>34</sub> MAM G<sub>35</sub>  
MA(MA)x(6)MAM G<sub>36</sub> MA(MA)x(44)MAM G<sub>37</sub> MA(MA)x(10)MAM G<sub>38</sub> MA(MA)x(1)MAM G<sub>39</sub> MA(MA)x(3)MAM G<sub>40</sub>  
MA(MA)x(9)MAM G<sub>41</sub> MA(MA)x(19)MAM G<sub>42</sub> MA(MA)x(1)MAM G<sub>43</sub> MAM G<sub>44</sub> MA(MA)x(3)MAM G<sub>45</sub>  
MA(MA)x(42)MAM G<sub>46</sub> MA(MA)x(31)MAM G<sub>47</sub> MA(MA)x(10)MAM G<sub>48</sub> MA(MA)x(1)MAM G<sub>49</sub> M G<sub>50</sub> MAMAM G<sub>51</sub> M  
G<sub>52</sub> MAMAM G<sub>53</sub> M G<sub>54</sub> M G<sub>55</sub> M G<sub>56</sub> M G<sub>57</sub> M G<sub>58</sub>

MA(MA)x(6)MAM G<sub>36</sub> MA(MA)x(44)MAM G<sub>37</sub> MA(MA)x(10)MAM G<sub>38</sub> MA(MA)x(1)MAM G<sub>39</sub> MA(MA)x(3)MAM G<sub>40</sub>  
MA(MA)x(9)MAM G<sub>41</sub> MA(MA)x(19)MAM G<sub>42</sub> MA(MA)x(1)MAM G<sub>43</sub> MAM G<sub>44</sub> MA(MA)x(3)MAM G<sub>45</sub>  
MA(MA)x(42)MAM G<sub>46</sub> MA(MA)x(31)MAM G<sub>47</sub> MA(MA)x(10)MAM G<sub>48</sub> MA(MA)x(1)MAM G<sub>49</sub> M G<sub>50</sub> MAMAM G<sub>51</sub> M  
G<sub>52</sub> MAMAM G<sub>53</sub> M G<sub>54</sub> M G<sub>55</sub> M G<sub>56</sub> M G<sub>57</sub> M G<sub>58</sub>

Startpositions: 198825 202971

Seq 1: **ttctgagttctt**gtgtccacttcttatagcctgt-ctgtcccttt-ccttgctactctgggatcaacagtcacctcttgaacttttggggtgcttgcaa 100  
Seq 2: .....t...gg....-c.tt.t.c....-t...c.....t.t....g....gg.....g.....g....caacag..ca.c...

Seq 1: ---tagttcctct-ccctactcctaatttacggca-ggcc-ttagaaaccataaccttatttttaaaggtgtaaaaaaaaaaagatttaagacaaaagcaa 200  
Seq 2: gcc...a....-c...g.t.t....c....--.g..aat.....g.....--...a.-...acc..g...gc--...-c.t.g..gg...c.

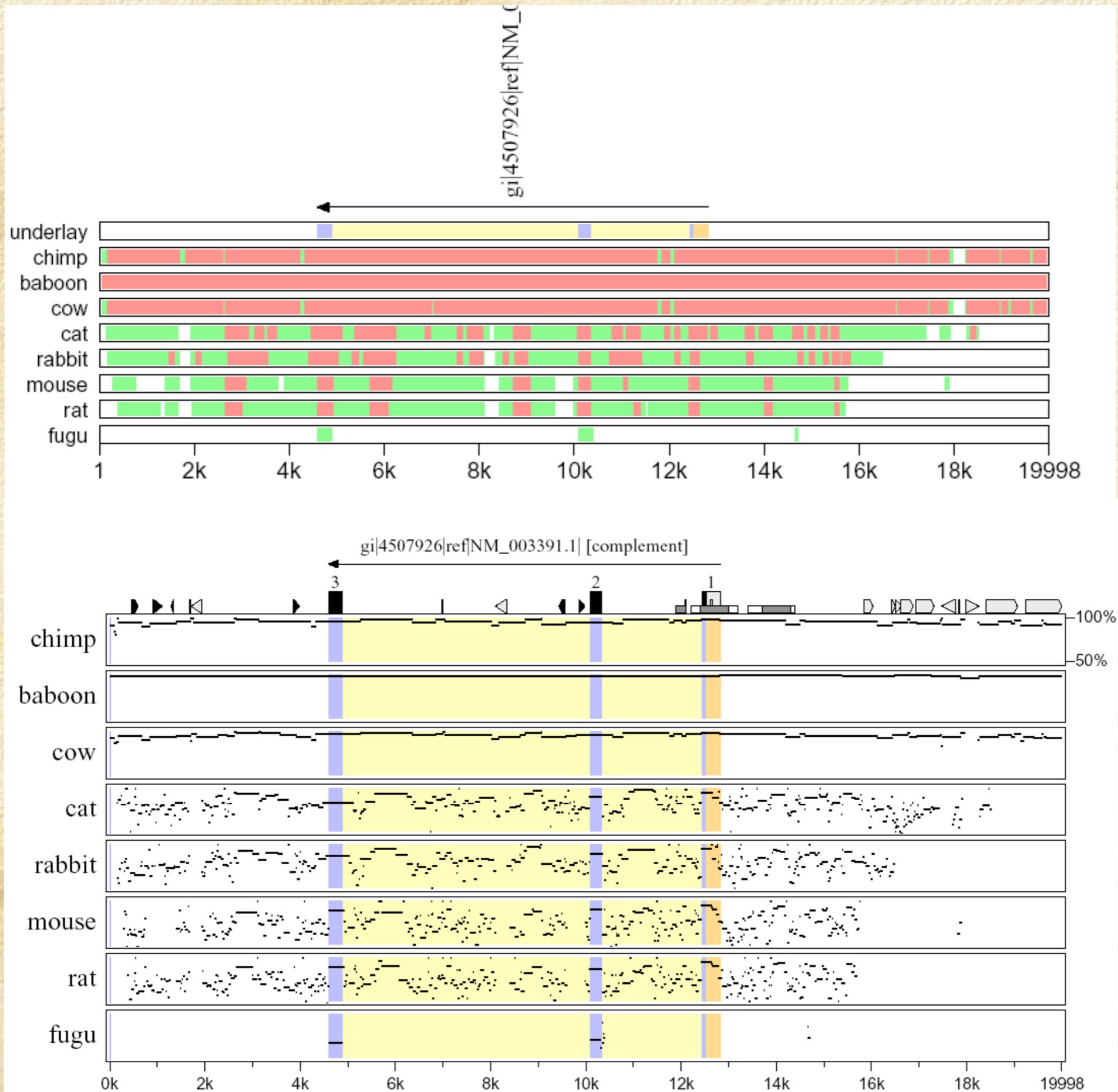
Seq 1: ggggctt-gg-gtgctttccttatgga--ct-t--aggcctggtaacatct-gttctggccacttagaggccttgtgtgctatttcttgttttcaggtgc 300  
Seq 2: ...t..cg..c....--....gca...cg..g.ac.....a.c..g.-.t.g.....tg.-.....-a.-...-g.--c..g.c-....gaa

Seq 1: gttttgcaggaggggacgttg-ttgagttccaaacaggtgaggtattgcac--actagcaaacacatgagaagaaggcggaggaattgggagaaaaataa 400  
Seq 2: -. .g.g.....t.c.g....ca.-.c...ct.....c.-....a.g.ag...g.gg.gt--...t.-....-.t..a.t.ca..c..g.....c.

Seq 1: aaagaatgcagcaggccaggttag-caggaacgttaagacgggtga-cggagaacagcaaagcctggaagcaagccgccgtggagaaggaa--g---aact 500  
Seq 2: .....cag..a.a..gg.....cc.-....accg....caca.ga..t.-.gg.....t..ctc.g...tt...aa.-.....ggg.aca....

Seq 1: gtgctgaggtgagttgctgtgacaaccaggctgattttgagtatgtaaacaccaaaccttgttcttggctgccgctcagctcagcgggctttggagcct 600  
Seq 2: .....-.....c...cc..gg..g.....tg.....c.....a..t..c.....a.....g.c.....

# MultiPipMaker - output



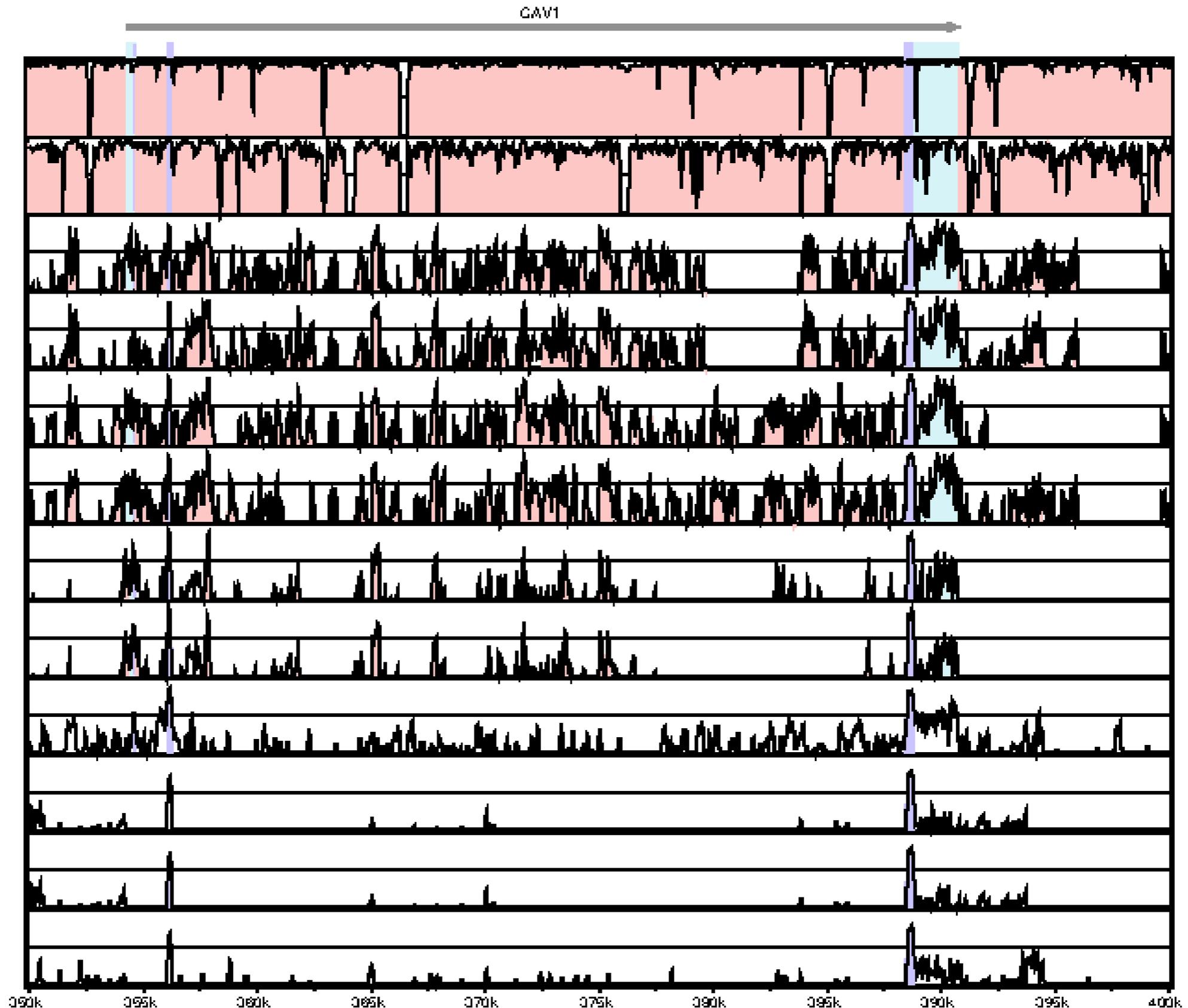
# MAVID/VISTA - output

## MAVID zoo

- Alignment 1  
Seqs: human/chimp  
Criteria: 75%, 100 bp  
Regions: 16
- Alignment 2  
Seqs: human/boon  
Criteria: 75%, 100 bp  
Regions: 26
- Alignment 3  
Seqs: human/cat  
Criteria: 75%, 100 bp  
Regions: 82
- Alignment 4  
Seqs: human/dog  
Criteria: 75%, 100 bp  
Regions: 62
- Alignment 5  
Seqs: human/cow  
Criteria: 75%, 100 bp  
Regions: 73
- Alignment 6  
Seqs: human/pig  
Criteria: 75%, 100 bp  
Regions: 73
- Alignment 7  
Seqs: human/mouse  
Criteria: 75%, 100 bp  
Regions: 21
- Alignment 8  
Seqs: human/rat  
Criteria: 75%, 100 bp  
Regions: 19
- Alignment 9  
Seqs: human/chicken  
Criteria: 75%, 100 bp  
Regions: 3
- Alignment 10  
Seqs: human/fugu  
Criteria: 75%, 100 bp  
Regions: 1
- Alignment 11  
Seqs: human/tetra  
Criteria: 75%, 100 bp  
Regions: 2
- Alignment 12  
Seqs: human/zebrafish  
Criteria: 75%, 100 bp  
Regions: 1

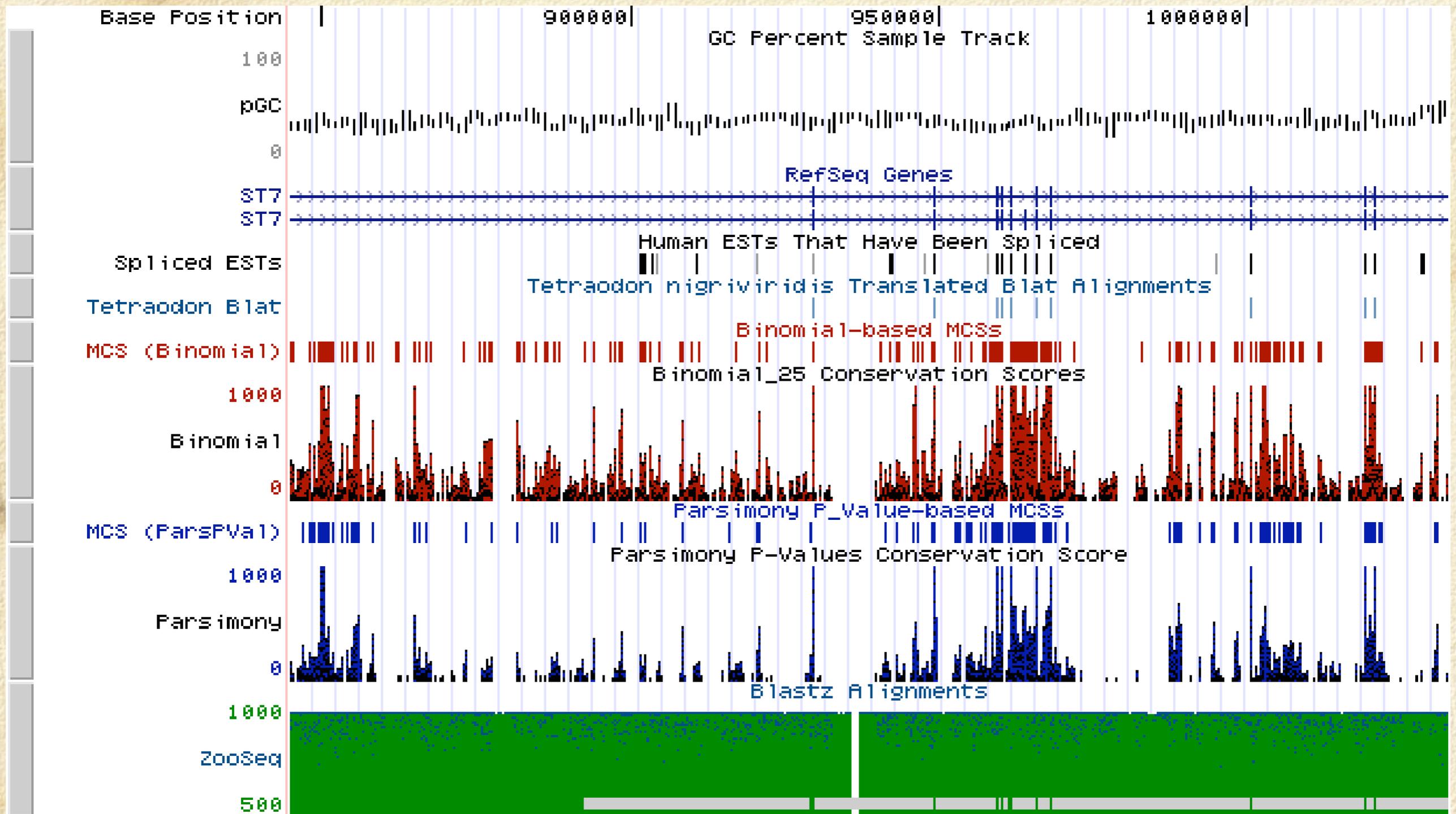
X-axis: human  
Resolution: 25  
Window size: 100 bp

- gene
- exon
- UTR
- CNS



# Genomic Targets for Comparative Sequencing

<http://genome.ucsc.edu/>



# References

---

- **BAlIbASE: a benchmark alignment database for the evaluation of multiple alignment programs**

*Thompson JD, Plewniak F, Poch O. Bioinformatics. 1999 Jan;15(1):87-8.*

- **A comprehensive comparison of multiple sequence alignment programs**

*JD Thompson, F Plewniak, and O Poch Nucleic Acids Res. 1999 27: 2682-2690.*

- **Quality assessment of multiple alignment programs**

*FEBS Letters Volume 529, Issue 1 , T. Lassmann and E Sonnhammer 2 October 2002, Pages 126-130*

- **Recent progress in multiple sequence alignment: a survey.**

*Notredame C. Pharmacogenomics. 2002 Jan;3(1):131-44. Review.*

- **Strategies for multiple sequences alignment**

*HB Nicholas Jr, AJ Ropelewski and DW Deerfield II, BioTechniques 32:572-591*

# This talk URLs

---

- <http://genome.nci.nih.gov/talks/msa.html>
- <http://helix.nih.gov/talks/>